

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

TVISION INSIGHTS, INC.,
Petitioner,

v.

THE NIELSEN COMPANY (US), LLC,
Patent Owner.

Inter Partes Review Case No. IPR2023-01014
U.S. Patent No. 11,470,243

SUPPLEMENTAL DECLARATION OF DR. ELI SABER

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
II.	OPINIONS ON CHALLENGED GROUND 1.....	3
A.	OPINIONS REGARDING TIAN’S PURPORTED TEACHING OF “REDUC[ING] A RESOLUTION”	3
1.	<i>Tian’s System Does Not Include Instructions, a Device, or a Process for Reducing the Resolution of An Image.....</i>	4
2.	<i>Applying Tian’s teachings would not result in a useable “audience measurement system,” as claimed</i>	10
B.	OPINIONS REGARDING PETITIONER’S MAPPING OF THE CLAIMED TWO-FRAME PROCESS	19
C.	OPINIONS REGARDING PETITIONER’S MOTIVATION TO COMBINE TIAN WITH LU-577/LU-000/LERT	23
1.	<i>Modifying Lu with Tian’s Teachings does not achieve Petitioner’s proposed benefit, and any benefit would be outweighed by the modification’s drawbacks.....</i>	24
a)	The Proposed Modification to Process Reduced- Resolution Images Would Not Reduce the Net Processing Resources Required.....	25
b)	The Proposed Modification to Perform Head Orientation Determination Is Flawed.....	43
c)	Petitioner’s Proposed Modification to Rely on Head Tracking Would Not Reduce the Net Processing Resources Required and Would Decrease the Performance of Face Recognition.....	49
III.	OPINIONS REGARDING THE ‘243 PATENT’S FIELD OF ENDEAVOR.....	54
IV.	CONCLUSION.....	56

I, Dr. Eli Saber, hereby declare the following:

I. INTRODUCTION

1. I have been asked to additionally respond to certain arguments raised by Petitioner in Petitioner's Petition ("Petition"). All of my opinions expressed in my original declaration (Ex. 2001) remain the same. In addition to the materials I reviewed in preparing the original declaration (see *Declaration of Dr. Saber* (Ex. 2001)¹, Section B), I also reviewed the below materials:

- Deposition transcript of Dr. David Scott Doermann (Ex. 2018);
- *Machine Analysis of Facial Behavior: Naturalistic and Dynamic Behavior*, Maja Pantic, The Royal Society, 2009 (Ex. 2021);
- *Super-Resolution Reconstruction and Its Future Research Direction*, Patanavijit, 2009 (Ex. 2022);
- *Discrete-time Signal Processing*, Oppenheim et al., Third Edition, Pearson, 2010 (Ex. 2014);
- *Analysis and Comparison of Various Image Downsampling and Upsampling Methods*, Abdou Youssef, Department of EECS, December 1998 (Ex. 2023);

¹ Citations to *Dec.* are to paragraphs of Ex. 2001, Declaration of Dr. Eli Saber.

- *1/2.5 Inch 5Mp CMOS Digital Image Sensor*, Aptina Imaging (Ex. 2024);
- *Aliasing*, Bruno Olshausen, October 10, 2000 (Ex. 2025);
- U.S. Patent No. 7,697,012 to Lee et al. ("*Lee '012*") (Ex. 2026);
- U.S. Patent No. 7,298,866 to Lee ("*Lee '866*") (Ex. 2027);
- U.S. Patent Application Publication No. 2012/0281028 to Orlick et al. ("*Orlick*") (Ex. 2028);
- *Quadrature Prefiltering for High Quality Antialiasing*, Guenter et al., Georgia Institute of Technology, October 1996 (Ex. 2029);
- European Patent No. 0694188 to Arbeiter et al. ("*Arbeiter*") (Ex. 2030);
- *Filtering by Repeated Integration*, Paul Heckbert, ACM, Vol. 20, No. 4, 1986 (Ex. 2031);
- *Using the Convolution Filter Process*, Release 9.2, April 14, 2008 (Ex. 2032);
- *Convolution Algorithms*, Sidney Burrus, April 10, 2010 (Ex. 2033);
- *Face Recognition Algorithms*, Universidad del Pais Vasco, June 16, 2010 (Ex. 2034);
- *Facial Feature Based Head Tracking and Pose Estimation*, Jari Hannuksela, Department of Electrical and Information Engineering, University of Oulu, 2003 (Ex. 2035);

- *Is Your TV Too Big for Your Room?*, Ann Porter, October 3, 2007 (Ex. 2036);
- *Sectional Sofa With Two Pillows*, https://ilovebbcsmlife/product_details/83389857.html (Ex. 2037);
- *Face Detection and Tracking Techniques: A Review*, Intl. Journal of Advanced Research in Computer Science, Vol. 2, No. 5, October 2011 (Ex. 2038); and
- Any other materials I cite in support of this Declaration.

II. OPINIONS ON CHALLENGED GROUND 1

A. Opinions Regarding Tian’s Purported Teaching of “Reduc[ing] a Resolution”

2. Petitioner relies on *Tian* for teaching *reduc[ing] a resolution*², as recited in dependent claims 4, 11, and 18³. *Petition*, 26-27. Namely, Petitioner maps *Tian*’s investigation of how “different image resolutions” affect each step in facial expression analysis to the *reduc[ing] a resolution...* limitation. *Id.* To explore various resolutions, low-resolution images are “down-sampled” from original

² Claim language is underlined and italicized for ease of reference.

³ Petitioner applies the same mapping and arguments used in dependent claim 4 for similar dependent claims 11 and 18.

images. *Tian*, Abstract. Petitioner argues the down-sampling teaches reduc[ing] a resolution.... *Petition*, 26; *see also*, *Dec.*, ¶ 45.

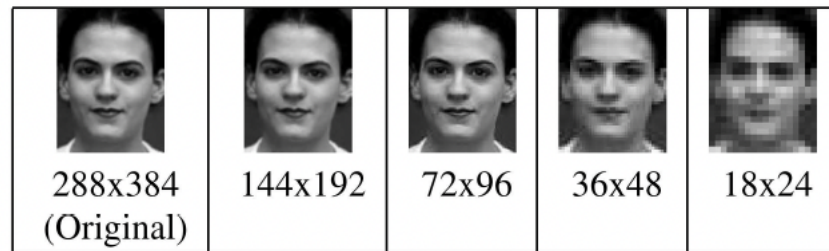
3. Each of claims 4, 11, and 18 recite an **active step** of either processor circuitry executing machine readable instructions to...reduce a resolution of a first image, an audience measurement device to ... reduce a resolution of the first image, or a method step of reducing a resolution of the first image. '234 Patent, 15:56-58, 16:49-51, 17:41-43. As I previously discussed, it is my opinion *Tian*'s passive disclosure of down-sampled images being contained in a pre-existing database **does not teach** the required active resolution reduction. *Dec.*, ¶¶ 46-48. In other words, *Tian* does not teach instructions, a device, or a process where images are actually down-sampled as part of the facial expression analysis system, as further explained below.

1. Tian's System Does Not Include Instructions, a Device, or a Process for Reducing the Resolution of An Image

4. To explore the "effects of different image resolutions" on facial expression analysis, *Tian* uses images "in five different resolutions from Cohn-Kanade database." *Tian*, 000002 (Table 1 caption). *Tian* specifies the lower resolution images are down sampled from original images to clarify that various resolutions of the same image are analyzed. *Dec.*, ¶ 46.

5. Petitioner applies the "down-sampled" language to support that "*Tian creates* lower-resolution images by down-sampling from original images." *Petition*,

26. Contrary to Petitioner’s assertion, *Tian* does not teach any instructions, device, or process for actively creating lower-resolution images. Instead, a POSITA would have understood *Tian* uses a set of pre-reduced images to test a variety of algorithms used in facial expression analysis to investigate their performance at each of five different resolutions. For example, Table 1 expressly discloses an exemplary image set of “five different resolutions **from Cohn-Kanade database**” that are investigated:



Tian, 000002. Thus, it is my opinion that a POSITA would read the caption of Table 1 as indicating all images, including the lower resolution images, come “from the Cohn-Kanade database.” *See also, Dec.*, ¶ 49.

6. Additionally, *Tian* summarizes the results of the investigation of the five different resolutions in Table 3, reproduced below:

Face Process		 288x384 (Original)	 144x192	 72x96	 36x48	 18x24
Face Acquisition	FD	100%	100%	100%	100%	0%
	HPE	98.5%	98%	98.2%	97.8%	98%
Feature Extraction	G1	Yes	Yes	Yes	No	No
	G2	Yes	Yes	Yes	Yes	No
	AP	Yes	Yes	Yes	Yes	Yes
FACS AUs	G1	90%	90.2%	89.8%	N/A	N/A
	G2	71%	70.8%	72%	54.3%	N/A
	AP	90.7%	90.2%	89.6%	72.6%	58.2%
	G1+AP	92.8%	93%	92.2%	N/A	N/A
	G2+AP	91.2%	90.8%	90%	87.7%	N/A
Basic Expressions	G1	92.5%	91.8%	91.6%	N/A	N/A
	G2	74%	73.8%	72.9%	61.3%	N/A
	AP	91.7%	92.2%	91.6%	77.6%	68.2%
	G1+AP	93.8%	94%	93.5%	N/A	N/A
	G2+AP	93.2%	93%	92.8%	89%	N/A

Tian, 000005.

7. Various different facial expression analysis processes are performed on each image resolution, including face detection, head pose estimation, and feature extraction. *Id.* Indeed, three different algorithms are tested for feature extraction; namely, (1) extracting geometric features via “feature tracking,” (2) extracting geometric features via “feature detection,” and (3) extracting features via Gabor wavelets. *Id.* Because multiple different algorithms are applied to each resolution by *Tian*—including different algorithms for performing the same facial expression analysis step—a POSITA would understand *Tian*’s express disclosure of using “five

different resolutions from Cohn-Kanade database” indicates the reduced images come from the database. *Tian*, 000002 (Table 1 caption). A POSITA would further understand that image resolution would not be reduced as a part of the image analysis process. A POSITA would have understood that repeated resolution reduction would waste resources, and thus would have read *Tian* to use images at the desired resolutions “from” the database. *Tian*, Table 1.

8. To further indicate *Tian* does not employ any instructions, device, or process for reduc[ing] the resolution of a first image, *Tian* investigates the effects of various image resolutions on common “automatic facial expression analysis (AFEA) systems.” *Tian*, Abstract. These systems are developed for recognizing facial expressions in “very high resolution frontal face[]” images. *Id.* *Tian* explores the effects of various real-world image resolutions within known AFEA systems. *Id.* Figure 1 below outlines an exemplary AFEA system being tested to investigate the effects of image resolutions:

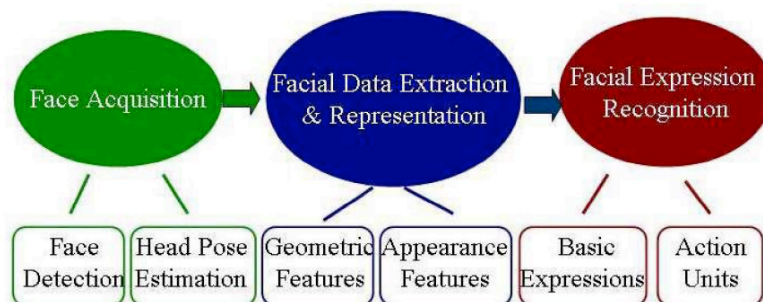


Figure 1: Basic Structure of Facial Expression Analysis Systems.

Tian, 000001. This is consistent with, and confirmed by, Table 3 (*Tian*, 000005, reproduced above), which shows each of these steps being performed for each of the five resolutions coming “from” the database in showing that resolution reduction would not be a part of *Tian*’s AFEA system, and would not be performed by *Tian*. *Tian*, 000002 (Table 1 caption).

9. The exemplary AFEA system structure outlined in Figure 1 does not include a step of reduc[ing] the resolution of a first image. Rather, all of the above elements of the system are applied to an image regardless of what resolution the image is at. In other words, the common processes of face detection, head pose estimation, feature extraction, and expression recognition do not involve any resolution reduction step. *Machine Analysis of Facial Behavior: Naturalistic and Dynamic Behavior*, Maja Pantic, The Royal Society, 2009 (Ex. 2021) at p. 3505 (outlining the “three steps” of facial expression recognition, none of which include reducing a resolution of the image). Rather, as *Tian* discloses, these systems are commonly developed to handle high-resolution images, so they would not contain a reduction resolution step. *Tian*, Abstract.

10. Additionally, a POSITA would not be motivated to add a step of resolution reduction into the AFEA systems being tested by *Tian*, as *Tian* himself disparages using images of lower resolution in recognizing facial expressions. *Tian* discloses “low resolution input images” are “among the factors that complicate facial

expression analysis.” *Tian*, 000006. Indeed, researchers developed techniques to combat this known disadvantage, such as employing super-resolution in AFEA systems to “get higher resolution images.” *Tian*, 000006; *see also*, *Super-Resolution Reconstruction and Its Future Research Direction*, Patanavijit, 2009, (Ex. 2022) pp. 156 (describing research efforts applying super-resolution processes to recognize faces in images and applying such processes to “multi-model cases such as smiling or angry images”). Thus, because lower-resolution images were known to further complicate expression analysis for AFEA systems, a POSITA would not have been motivated to include instructions, a device, or a process step that adds this complication into the AFEA systems tested by *Tian*.

11. *Tian* is not looking to purposely reduce resolutions of images applied to AFEA systems, but understands that low-resolution images will naturally occur in “real environments.” *Tian*, Abstract. Rather, *Tian* indicates the images processed in AFEA systems are ideally of high resolution, as lower resolution “input images” complicate AFEA and make AFEA “**more difficult**.” *Tian*, Abstract, 000002; *see also*, *Dec.*, ¶ 121. Further, *Tian*’s discussion of techniques to increase image resolution (e.g., super-resolution) further indicates reducing a resolution of an image should be avoided. Additionally, if a low-resolution image is unavoidable, *Tian* understands the importance of mitigating the corresponding AFEA complications by increasing the resolution, when possible. *Dec.* ¶¶ 121-122. Thus, a POSITA would

not look at *Tian* and understand *Tian* to teach an AFEA system performing instructions, including a device, or performing a process of reduc[ing] a resolution of a captured image.

2. *Applying Tian’s teachings would not result in a useable “audience measurement system,” as claimed*

12. The details of the resolution reduction process applied in the ’243 Patent are relevant to how the audience measurement system operates. Namely, the ’243 Patent describes applying “pixel binning” to images captured “without use of the illumination source.” ’243 Patent, 3:50-56. The binning process sums certain neighboring pixels to increase the contrast (i.e., increase the light per pixel), while decreasing the resolution of the image. ’243 Patent, 3:54-56, 11:10-19. For this reason, images taken without the illumination source that are “likely to be of low contrast” are processed via pixel binning to increase the contrast. ’243 Patent, 11:5-10. Most images in the audience measurement system will be taken in the “majority mode,” without the illumination source being on, in order to decrease audience annoyance caused by the illumination source. ’243 Patent, 8:52-54, 8:60-64, 10:15-17 (describing the illuminator is not illuminated in the “majority mode” of image capture), 3:12-24 (describing the illumination source as annoying audience members when the light is turned on). Thus, the pixel binning procedure provides the advantage of increasing the contrast in an image captured in a low-light environment. This advantage is specific to audience measurement systems where

images of the audience are taken in low-lit rooms while people are watching TV.

'243 *Patent*, 10:58-63 (describing unilluminated images are too dark and low-contrast for analysis).

13. Additionally, the increased contrast provided by the pixel binner allows the system to “distinguish between shapes in the frames of image data,” such as distinguishing a head of a person. '243 *Patent*, 11:31-34. The higher-contrast images are stored in the frame cache. '243 *Patent*, 11:34-36. The system references these “**recently-captured**” images in the frame cache to determine when to turn the illumination source on for a resolution-demanding process, such as facial recognition. '243 *Patent*, 9:7-14, 9:26-31, 9:33-37 (describing the system placing the sensor into “minority mode” for illuminating “**succeeding**” images that are indicated for facial recognition). Specifically, a hi-res trigger analyzes the frame cache images to “determine” whether a head is in the image, and what orientation the head is in. '243 *Patent*, 9:14-22. The increased contrast also places the recently captured frame cache images in the proper condition for head detection and orientation determination operations. Thus, the pixel binning procedure also provides the advantage of enhancing the low-contrast images to a higher contrast for immediate analysis to determine whether succeeding frames require illumination. Furthermore, the pixel binner is disposed within the image sensor. '243 *Patent*, Fig. 3, 8:44-50. Thus, the resolution reduction process is performed by the image sensor

component, indicating the process is performed quickly, immediately after images are captured. This is so the preceding frames are appropriately processed for the head detection analysis to be relevant in determining whether the illuminator should be activated for the next frame capture.

14. Based on the important outcomes of the resolution reduction process (i.e., higher contrast images and fast analysis) of the '243 Patent, a POSITA would understand a combined system that does not reduce the resolution appropriately will not be able to designate frames for illumination and facial recognition, thereby rendering the system useless for audience measurement. A POSITA, only given *Tian's* limited details regarding reduced-resolution images produced not in real time via down-sampling, would understand that a useable audience measurement system would not result from applying *Tian's* teachings.

15. As discussed above, applying the '243 Patent's resolution reduction process of pixel binning allows the system to detect heads and identify audiences in real time or near real time. *See* ¶ 13. Contrastingly, *Tian's* reduction resolution process is **not** a real-time process. As discussed above, *Tian* does not teach actively reducing the resolution of captured images but rather uses reduced images “from” a database. *See*, ¶¶ 4-7. Because *Tian's* pre-processing for resolution reduction occurs offline rather in real-time (i.e., occurs before *any* analysis begins), all of the “succeeding” frames would already be captured in the audience measurement system

before the preceding frame would be analyzed to trigger illumination. Thus, in a system applying *Tian*'s pre-processing of resolution, no frames would be illuminated because the analysis would not occur in time for the triggering of the illumination source, and accordingly the system would never successfully perform facial analysis. '243 *Patent*, 10:44-54 (images suitable for facial recognition are high-resolution images that are well-lit). Thus, such a system would be unable to identify audience members and useless as an audience measurement system or method.

16. Additionally, as discussed above, the resolution reduction process (i.e., pixel binning) applied in the audience measurement system of the '243 *Patent* increases the image contrast so that the unilluminated images may be analyzed to determine whether to trigger illumination for succeeding frames. This is important, as images captured in a low-lit environment, such as a media exposure environment, have low contrast. The pixel binning procedure increases contrast because neighboring pixels are summed together, such that "more light" is assigned to each block of pixels. '243 *Patent*, 11:10-18. *Tian*'s pre-processing of images to reduce the resolution is accomplished via down-sampling. *Tian*, 000002. Down-sampling **does not** increase the contrast of images. Down-sampling is a broad term that is generally defined as the "operation of reducing the sampling rate (including any prefiltering)." *Discrete-time Signal Processing*, Oppenheim et al., Third Edition, Pearson, 2010 (Ex. 2014) at p. 180. Exemplary well-known down-sampling methods

include “subsampling” (i.e., decimation) and “various forms of interpolation.” Ex. 1003, ¶ 15; *Analysis and Comparison of Various Image Downsampling and Upsampling Methods*, Abdou Youssef, Department of EECS, December 1998 (Ex. 2023) at p. 2 (describing “standard methods” for “downsampling” include “decimation” (i.e., subsampling) and “bilinear interpolation.”).

17. Decimation (including any prefiltering) reduces the resolution of the image by removing certain pixels, as the sampling rate is reduced by a factor of “M.” *Discrete-time Signal Processing*, Oppenheim et al., Third Edition, Pearson, 2010 (Ex. 2014) at pp. 181-184. For example, to reduce the resolution by a factor of two, a decimation algorithm would apply a factor of $M=2$ to the image, which would remove half of the pixels from the rows and columns of the image, yielding an overall image that is a quarter of the original image. Merely removing pixels will not affect the contrast of the image, but will create aliasing artifacts for frequency components of the image that are below the Nyquist frequency, as shown below (middle image). Techniques such as bilinear interpolation reduce the resolution of the image by averaging neighboring pixels together and replacing the pixels with the “average pixel value.” *Analysis and Comparison of Various Image Downsampling and Upsampling Methods*, Abdou Youssef, Department of EECS, December 1998 (Ex. 2023) at p. 2. Because pixel values are averaged, the resulting image is blurrier than the original (left image), as shown below (right image):



Analysis and Comparison of Various Image Downsampling and Upsampling Methods (Ex. 2023), at FIGs. 5-6. As can be seen from the images above, neither decimation nor bilinear interpolation increase contrast.

18. Unlike down-sampling, pixel binning increases contrast by combining adjacent pixels by **summing** the pixels together to “improve[] low-light performance.” *1/2.5 Inch 5Mp CMOS Digital Image Sensor*, Aptina Imaging (Ex. 2024) at p. 27. Thus, because the Cohn-Kanade database images are reduced by down-sampling, applying *Tian’s* teachings would create an image having both low resolution **and** low-contrast. A low-contrast image is not suitable for analysis in the audience measurement system of the ’243 Patent, as the “shape” of a head would not be reliably detected for determining when to turn the illumination source on for subsequent images. ’243 Patent, 11:31-34 (describing the enhanced “higher contrast levels” provided by the pixel binning enabling the system to “distinguish between shapes” in the image), 9:14-31 (describing the system detecting a head and the orientation of a head to determine whether to trigger illumination for succeeding

frames that should undergo facial recognition). Furthermore, even high-resolution images, when captured without the illuminator, would suffer from low contrast and thus not be suitable for analysis in the audience measurement system. '243 Patent, 10:58-67. Thus, because *Tian's* teachings would produce a low-contrast image, a POSITA would understand any combination with *Tian's* teachings to reduce a resolution would not make a useable audience measurement system.

19. Petitioner specifically points to “subsampling” and “interpolation” as techniques known to reduce image resolution. Petition at p. 2; *see also* Ex. 1003, ¶ 15. However, as discussed above, neither of these techniques increase the contrast of the image, and thus applying these techniques to an audience measurement system would render the system useless for detecting audience members and identifying them. *See* ¶¶ 16-17.

20. Further, *Tian's* disclosure of using images that are only down-sampled, without any other disclosed pre-processing steps, would not result in a useable audience measurement system, because the down-sampled image would not be ready for analysis. Down-sampling is vulnerable to artifacts resulting from aliasing. Aliasing occurs when the sampling rate is too low (i.e., not “at least twice the highest frequency” in the input image). *Aliasing*, Bruno Olshausen, October 10, 2000 (Ex. 2025) at p. 1. Because down-sampling reduces the sampling rate, down-sampling without a prefilter leads to aliasing when the sampling rate is less than twice the

Nyquist frequency. *See, e.g., Lee '012* (Ex. 2026), 1:25-34; *Lee '866* (Ex. 2027) (describing filters that may be used to “remove” aliasing when the image is down-sampled); *Aliasing*, Bruno Olshausen, October 10, 2000 (Ex. 2025) at p. 1. If prefiltering techniques—such as antialiasing—are not applied, the aliasing in the down-sampled image will cause “visual artifacts” in the image, such as “jagged patterns” on the edges of objects that degrade the “quality of the image.” *Orlick* (Ex. 2028), [0030] (describing the “downsampling step” to result in “aliasing” that could “cause visual artifacts” in the image). These jagged patterns and visual artifacts distort the shape of the head, thereby interfering with head detection and head pose estimation.

21. Many prefiltering techniques exist and are employed to mitigate aliasing affects when down-sampling images, including applying low-pass filters. *See, e.g., Discrete-time Signal Processing*, Oppenheim et al., Third Edition, Pearson, 2010 (Ex. 2014) at p. 206; *Lee '012*, 1:33-40 (applying a low-pass filter to the signal to prevent aliasing prior to down-sampling the signal). These prefiltering techniques often result in blurring, since the prefiltering operation is designed to remove high frequency content from the signal in order to prevent aliasing. Prefiltering techniques may apply in the context of *Tian's* non-real-time down-sampling. However, these same prefiltering techniques may not be suitable in a system requiring real-time processing of images, such as an audience measurement system. Indeed, it was

known that certain “prefiltering methods are **slow**, costly, and **complex**.” *Quadrature Prefiltering for High Quality Antialiasing*, Guenter et al., Georgia Institute of Technology, October 1996 (Ex. 2029) at p. 334. For example, depending on the amount of down-sampling applied, the prefiltering algorithm may have to extend “over many samples” and require a “long filter response,” which increases the complexity of the filter. *Arbeiter* (Ex. 2030), [0004-0005].

22. If such a complex, slow, prefiltering algorithm is used to perform down-sampling in the proposed combination, a POSITA would understand the media exposure environment may change significantly by the time the modified system detects a head in the image, indicating a viewer is present. For example, the program present on the TV may change, or the person may have left the environment entirely before the system performs facial recognition to identify an audience member based on the detection of a head. In these scenarios, any audience member identification would not provide accurate viewing measurements.

23. Additionally, during a slow and complex prefiltering algorithm, an audience member may move their head. Thus, using the location of the head in the first image to perform facial recognition on a subsequent image of potentially higher resolution would fail. The claimed audience measurement system uses the location of the head in the first image when performing facial recognition on a subsequent image of potentially higher resolution. *See, e.g., '243 Patent*, claims 5, 6, 12, 14, and

19. Such an approach would not result in a viable audience measurement system when applying *Tian*'s down-sampling.

24. For the above reasons, it is my opinion *Tian*'s vague disclosure that images "are down-sampled" previously and then analyzed for facial expressions **does not teach or suggest** the claimed "machine-readable instructions to ... reduce a resolution of a first image," "an audience measurement device to ... reduce a resolution of the first image," or the method step of "reducing a resolution of the first image."

B. Opinions Regarding Petitioner's Mapping of the claimed Two-Frame Process

25. Dependent claims 5, 6, 12, 14, and 19 generally recite performing face recognition on a second image based on the location of the head determined in the first image. '243 Patent, 15:61-16:9, 16:54-61, 17:1-7, 17:40-47. Dependent claims 8, 13, and 20 depend from claims 6, 12, and 19, respectively, and therefore similarly incorporate the face recognition being performed on a different image (i.e., second image) from which the head was detected (i.e., first image), and orientation determined. '243 Patent, 16:15-20, 16:62-67, 17:58-61. For ease of reference, I refer to the above process as the two-frame process, because the process (1) identifies the location of a head in a first frame and (2) uses the identified head location from the first frame to perform facial recognition in a second frame.

26. Petitioner maps *Tian*'s "teach[ing]" of "determining a face/head location in a first image and using this location for subsequent images unless a change is detected" to the two-frame process. *Petition*, 28, 30, 37, 42. It is my opinion that *Tian* **does not teach** applying such a two-frame process, but rather implements a single-frame process for facial analysis, as discussed below.

27. *Tian* discloses:

Face acquisition is a processing stage to automatically find the face region for the input images or sequences. **It can be a face detector to detect a face in each frame or just detect face in the first frame and then track the face in the remainder of the video sequence.** In order to handle large head motion, head finding, head tracking and pose estimation can be applied to a facial expression analysis system.

Tian, 000001 (Petitioner uses **bold text** as support). *Tian*'s "track[ing]" of the face indicates the face still moves between frames, and thus the location of the face is updated in each frame and not based on an original location. Furthermore, *Tian*'s disclosure of a face detector detecting a face "in each frame" also indicates the location of the face is updated for each frame. Regardless of either method applied, *Tian* is locating a head and conducting facial analysis (e.g., expression recognition) **using a single frame** (i.e., a single-frame process), and thus does not perform a two-frame process (i.e., determining head/face location in first frame and using that location to perform facial recognition in the second frame), as claimed.

28. Additionally, the “head finding” and “head tracking” is applied during “face acquisition” in AFEA systems to “handle large head motion.” *Id.* Thus, not only must the face be tracked (i.e., located) in current frames to follow movement of the face, but the head is also tracked so the system may handle “head motion.” Thus, because the common AFEA systems being tested by *Tian* track both the head and the face between video frames to account for movement, the system determines and uses the location of the head and face **in the current frame** for expression recognition of that current frame, rather than the location from a previous frame. If the location in a previous frame would be used to successfully locate the head and face in a subsequent frame, then there would be no need for tracking between frames. Thus, the use of tracking indicates the location of the head and face is found in each frame, and not based on a previous frame’s head and face location.

29. Dr. Doermann attempts to support the assertion that *Tian* applies the same head location “unless a change is detected” by pointing to *Tian*’s disclosure of locating the “face” and “face features” in the “initial frame,” and then “**all face feature changes are automatically tracked in the image sequence.**” *Tian*, 000003; *Declaration of Dr. Doermann* (Ex. 1003), ¶ 62. However, tracking of the face feature changes indicates the location of the face and features would be detected in the current image for analysis of that image, rather than relying on an inaccurate location from a previous image. Thus, based on *Tian*’s disclosures, a POSITA would

understand *Tian* determines the location of the face and head for the image frame to be analyzed, rather than using a location detected in a previous image.

30. *Tian* further states that when analyzing low-resolution images, the images must be “well-aligned” to provide accurate results of expression recognition. *Tian*, 000006. The alignment of the faces within the images depends upon the tracking accuracy, and thus accurate analysis of low-resolution images requires accurately tracking the head and face. Thus, to ensure accurate face alignment, a POSITA would have understood *Tian* uses the head location detected *in the given frame* rather than the head and face location found in a previous frame.

31. Attempting to justify the statement that *Tian* uses the head location detected in a previous video frame for subsequent video frame analysis, Dr. Doermann opines that “video cameras were able to take multiple images in a relatively short period of time, in which a person’s head or face is unlikely to move significantly.” Ex. 1003, ¶ 64. However, this contradicts *Tian*’s relied-upon teachings that determine the head/face location in each frame in order to ensure the faces are “well-aligned.” *Tian*, 000001, 000006. *Tian* teaches that head/face tracking occurs for each frame and thus determines the location of the head regardless of the level of movement.

C. Opinions Regarding Petitioner’s Motivation to Combine Tian with Lu-577/Lu-000/Lert

32. As discussed above, the ’243 Patent discloses actively reducing the resolution of images via pixel binning to increase the contrast of images captured in unilluminated media exposure environments. *See* ¶ 13. Thus, the audience measurement system described in the ’243 Patent intentionally reduces the resolution in order to increase the image contrast. In the general “facial analysis” processes of *Tian* (*Petition*, 9), intentionally reducing the resolution of images would hinder the recognition performance. When viewing *Tian*’s results, the highest image resolution provides the best accuracy for both face detection and head pose estimation:

Face Process						
		288x384 (Original)	144x192	72x96	36x48	18x24
Face	FD	100%	100%	100%	100%	0%
Acquisition	HPE	98.5%	98%	98.2%	97.8%	98%

Tian, Table 3 (annotated, green denotes the highest resolution and red denotes the “head pose estimation” (HPE) and “face detection” (FD) accuracy result at the highest resolution); 000005.

33. The results indicate that there is a discrepancy between using the lowest resolution images and the highest resolution images. For example, a higher accuracy

results from performing head pose estimation on the highest image resolution. *Tian* understood this discrepancy, as most AFEA systems are developed for “very high resolution faces.” *Tian*, 000002.

34. As discussed below, Petitioner’s modification of *Lu-577/Lu-000/Lert’s* audience measurement system with *Tian’s* teachings would not have been motivated by Petitioner’s proposed benefits, and thus a POSITA would not have found it obvious to create the combined system.

1. Modifying Lu with Tian’s Teachings does not achieve Petitioner’s proposed benefit, and any benefit would be outweighed by the modification’s drawbacks

35. Petitioner proposes several modifications to Lu⁴ (i.e., *Lu-577/Lu-000/Lert*) applying *Tian’s* teachings to render the Challenged Claims obvious in Ground 1, including providing various benefits to motivate such modifications. *Petition*, 9-11. Petitioner’s benefits and modifications are outlined below:

- Modifying Lu to include *Tian’s* teaching of detecting a head and determining head orientation on reduced-resolution images for claims 4, 11, and 18 provides the benefit of reduced processing resources required. *Petition*, 10.

⁴ “Lu” refers to the Lu-577/Lu-000/Lert system herein.

- Modifying Lu to include *Tian's* head orientation determination performance for claims 4, 11, and 18 improves the performance of facial recognition by determining which images are “suitable.” *Id.*
- Modifying Lu to include *Tian's* head tracking in a video sequence for claims 5, 6, 12, 14, and 19 would save processing resources by reducing the number of times face detection is applied to images, as the face would be detected in the first image and that location applied in subsequent images. *Id.* at 9.

36. As I discuss below, it is my opinion that each of the above modifications involves drawbacks that would dissuade a POSITA from making such modifications. Further, these drawbacks would outweigh the proposed benefits resulting in a net increase in time and processing requirements.

a) The Proposed Modification to Process Reduced-Resolution Images Would Not Reduce the Net Processing Resources Required

37. Petitioner modifies Lu to include *Tian's* performing of head orientation determination on reduced-resolution images to render obvious claims 4, 11, and 18. *Petition*, 26-27, 36-37, 42. Petitioner argues that a POSITA would have been motivated to make such a modification to “reduce the time and resource consumption of tasks such as face detection and pose estimation, without substantially reducing the accuracy of these tasks.” *Petition*, 10 (citing Ex. 1003,

¶¶69-70). However, as discussed below, Petitioner’s modification **would not** save time and processing resources and **would** reduce the accuracy of face detection and pose estimation, as the disadvantages associated with the modification outweigh the purported benefits.

i. Tian’s Resolution Reduction Requires Additional Processing Resources

38. Neither Petitioner nor Dr. Doermann addresses the **additional** “resource consumption” caused by adding a step of reducing the resolution of the image. As discussed above, image resolution reduction—such as by down-sampling, as *Tian* suggests—is known to cause aliasing, which creates various unwanted artifacts and distortions within the image. *See*, ¶¶ 17, 20. The image distortion may interfere with downstream image processing, such as face or head detection. *See*, ¶ 20. To avoid such image distortion caused by aliasing, prefiltering methods are applied prior to down-sampling. *See*, ¶ 21. The most common and “fundamental” filtering operation is convolution, which passes a weighting function (or kernel) over the input signal. *Filtering by Repeated Integration*, Paul Heckbert, ACM, Vol. 20, No. 4, 1986 (Ex. 2031), at p. 315. The kernel is a matrix of a certain size that is successively applied to each pixel in the image (and its neighboring pixels) to generate a weighted average that replaces the value of the pixel being analyzed. *Using the Convolution Filter Process*, Release 9.2, April 14, 2008 (Ex. 2032); *Filtering by Repeated Integration*, Paul Heckbert, ACM, Vol. 20, No. 4, 1986 (Ex.

2031), at p. 315. “For a discrete signal of length M and a filter of width W , the number of operations required by direct convolution is in the order of $O(MW)$ ”.. *Filtering by Repeated Integration*, Paul Heckbert, ACM, Vol. 20, No. 4, 1986 (Ex. 2031), at p. 315. In other words, the number of operations is a multiple of the signal length (i.e., a multiple of the number of pixels).

39. Thus, in order to reduce the resolution of an image via *Tian*’s suggested down-sampling without introducing artifacts that negatively affect subsequent processes, a POSITA would understand that a pre-filtering process would be applied prior to down-sampling the image. This extra step would require additional computational resources, which likely outweigh any reduction in “resource consumption” from Petitioner’s proposed modification of performing head detection and orientation determination on a reduced-resolution image and potentially result in lower accuracy in performing head detection given the lower resolution image.

40. Dr. Doermann overstates the processing required for face detection and head orientation determination, opining that reducing the image resolution “can” reduce the “processing resources and time,” because “[s]ome approaches of face detection may require analyzing all pixels of an image to search for a face.” Ex. 1003, ¶ 69. Essentially, Dr. Doermann argues because the reduced-resolution image would have fewer pixels, the time and processing required to detect a face would decrease. However, Dr. Doermann does not consider that the resolution reduction

process itself, excluding extra processing to avoid aliasing, would likely also require analyzing all image pixels. For example, down-sampling the image resolution via bilinear interpolation involves analyzing every pixel to replace a pre-defined pixel window “by its **average pixel value**.” *Analysis and Comparison of Various Image Downsampling and Upsampling Methods* (Ex. 2023) at p. 2. Thus, the resolution reduction process itself suffers as much or more of the computational cost that Dr. Doermann points to as a hindrance.

41. Additionally, as discussed previously, anti-aliasing algorithms (also referred to as prefiltering algorithms) are performed as part of reducing the resolution via down-sampling in order to avoid aliasing, which creates unwanted artifacts within the image. *See* ¶¶ 20, 38. These anti-aliasing algorithms typically utilize convolution. *See* ¶ 38. Convolution requires a number of processing steps that is a multiple of the number of pixels in the image (i.e. for a one-dimensional signal, the complexity is $O(MW)$). *Id.* The kernel matrices applied in convolution analyze the **neighboring pixels** of **each** original pixel when calculating the weighted average result. *Using the Convolution Filter Process*, Release 9.2, April 14, 2008 (Ex. 2032); *Filtering by Repeated Integration*, Paul Heckbert, ACM, Vol. 20, No. 4, 1986 (Ex. 2031), at p. 315. Regardless of the matrix size (e.g., 3x3 or above), each original pixel is processed multiple times during convolution, as each original pixel is a neighbor to more than one pixel.

42. Convolution algorithms are, in general, computationally complex and expensive (especially for large kernels), as researchers have attempted to develop filters to decrease the computational complexity and make the convolution algorithm run faster. *See, e.g., Convolution Algorithms*, Sidney Burrus, April 10, 2010 (Ex. 2033). However, even “fast” convolution algorithms still require multiple multiplication and summation operations to produce a single output value. *Id.* These multiple operations must be performed on several pixels (i.e., a “block”) in order to output “even one” value. *Convolution Algorithms*, Sidney Burrus, April 10, 2010 (Ex. 2033) at p. 1.

43. Additionally, as discussed above, down-sampling itself is commonly carried out either via subsampling (i.e., decimation) or by bilinear interpolation. *See* ¶¶ 16-17. In some instances, decimation may not provide a user with their desired resolution, because decimation only reduces the pixels by a factor of M (e.g., 2, 4). *Id.* To reduce the resolution by a non-integer factor (i.e., keep 3 out of 4 pixels), other down-sampling methods that utilize appropriate filtering and interpolation are generally used. *Analysis and Comparison of Various Image Downsampling and Upsampling Methods* (Ex. 2023), at p. 155. For example, to reduce the resolution by a non-integer factor, a combination of up-sampling (i.e., interpolation) and decimation is applied to the signal. *Discrete-time Signal Processing*, Oppenheim et al., Third Edition, Pearson, 2010 (Ex. 2014) at pp. 190-191. Thus, reducing the

resolution by a non-integer factor is more computationally expensive than reducing by an integer factor. Indeed, Interpolation itself also requires processing **each pixel in the image**. *Id.* at pp. 184-185. Thus, Petitioner's proposed combination of *Tian's* suggested down-sampling to reduce the resolution requires *additional* processing resources be implemented for audience measurement scenarios, when compared to performing all analysis steps on the full-resolution images. Considering these additional processing resources, a POSITA would not have been motivated to apply *Tian's* purported teaching of resolution reduction to the audience measurement system of Lu to save processing resources.

44. Furthermore, Dr. Doermann cites to Ex. 1013 to support the notion that “some approaches of face detection may require analyzing all pixels of an image,” and therefore may be resource intensive. Ex. 1003, ¶ 69. Ex. 1013 discusses specific face detection procedures, namely “view-based approaches,” as being “computationally rather expensive.” Ex. 1003, ¶ 69; Ex. 1013, 000001. However, other common face detection methods, such as template matching, knowledge-based methods, edge detection, and feature-based methods, among others, typically require lower processing resources when compared to view-based methods that require the use of multiple views. Ex. 1023, 000003-000006; *Face Recognition Algorithms*, Universidad del Pais Vasco (Ex. 2034) at pp. 12-13; Ex. 2012, 3-4. This is because these methods typically employ geometrical/structural aspects within the image to

detect features of a face when performing face detection, and thus need not analyze multiple views of the face. *Id.* Given the above, a POSITA would understand locating a face in a full-resolution image would not require noticeably higher processing resources or time than performing a resolution-reduction process followed by performing face detection on the resulting reduced image (with the added risk of reduced accuracy). Thus, the proposed benefit of decreasing processing resources is negligible, and a POSITA would understand the additional processing requirements to reduce the resolution would outweigh such a benefit. I note common head pose estimation methods are similarly simplistic and require relatively less processing resources (i.e., under Dr. Doermann’s definition of “resource intensive,” they do not need to analyze every pixel in the image). *See, e.g.,* Ex. 1020 (surveying various pose estimation techniques); *Facial Feature Based Head Tracking and Pose Estimation*, Jari Hannuksela, Department of Electrical and Information Engineering, University of Oulu, 2003 (Ex. 2035) at p. 25 (describing a “simple pose estimation procedure” algorithm that computes the 3D pose of an object “from three points if the geometry of the object is known” (i.e., ellipse shape of a head)). For example, geometric methods of performing head pose estimation merely locate facial features in the image and use these locations to determine pose “from their relative configuration.” Ex. 1020, 000003. Some template matching methods merely

compare an image to a set of exemplary images “labeled with a discrete pose” to find a matching head pose. *Id.*

ii. Petitioner’s Proposed Modifications are Redundant

45. As discussed above, Dr. Doermann opines reducing the resolution of images may reduce “processing resources,” by reducing the number of pixels that need to be processed for face detection, citing Ex. 1013 that discloses “**some** approaches of face detection” require analyzing all pixels. *See* ¶ 40. However, Dr. Doermann ignores the fact that other proposed modifications (i.e., head/face tracking as described below) to the system *already* reduce the number of pixels to be analyzed, making the proposed benefit of applying resolution reduction redundant within the modified system.

46. Specifically, Petitioner further modifies Lu to include *Tian’s* purported teaching of determining “a face/head location in a first image,” and using that location for processing of subsequent images via head tracking. *Petition*, 9; *See also*, ¶¶ 26. As proposed by Petitioner, this modification already avoids the need for analyzing all pixels of the image, as Dr. Doermann’s cited Ex. 1010 states that utilizing head tracking allows a face detector to only scan a “localized region” of pixels without needing to analyze all pixels in the image. Ex. 1003, ¶ 64 (citing Ex. 1010, [0003]). Thus, Petitioner’s own head-tracking modification and supporting

background references render the purported advantages of the resolution-reduction modification redundant.

47. This redundancy further weighs the disadvantages of Petitioner's proposed resolution-reduction combination higher than any proposed advantages. For the reasons discussed above, the proposed advantages of Petitioner's reduction-resolution modification (i.e., decrease processing resources) are small, if not non-existent considering the increased computation required to down-sample the images and the additional risk of decreased accuracy by operating on lower resolution images. *See* ¶¶ 37-43. Additionally, any advantages are made redundant by other modifications (such as head face/tracking) in Petitioner's proposed system, thus making the disadvantages outweigh the advantages. For these reasons, a POSITA would not have been motivated to modify Lu to reduce the resolution of images based on *Tian's* teachings.

48. I note *Tian* itself does not suggest a motivation or benefit of decreased processing resources. This is because *Tian* investigates how well facial expressions can be recognized in "real-life" instances, where the images are **captured at** "low resolutions[]." *Tian*, 000002. Therefore, *Tian* does not consider or address the processing resource requirement when reducing image resolution because the goal of the research is to determine how images already captured at the low resolutions can be used for facial expression recognition. Rather, a POSITA would understand

the benefits related to processing resources of reducing the resolution of a captured image prior to performing head detection and head orientation determination would be outweighed by the costs (i.e., processing resources, time, and accuracy) of the resolution reduction process in the scenarios of the '243 Patent.

iii. Petitioner's Accuracy Analysis is Factually Flawed

49. The only potential disadvantage Petitioner and Dr. Doermann address regarding the proposed modification is the effect on accuracy of face detection and head pose estimation performed on images of reduced resolution. *Petition*, 10; Ex. 1003, ¶ 69. Namely, Petitioner claims performing such operations on reduced-resolution images would not “substantially” reduce “the accuracy of these tasks.” *Petition*, 10. Petitioner and Dr. Doermann cite various references (Exs. 1013-1015) to support such a statement, including relying on *Tian's* experimental results. *Petition*, 10; Ex. 1003, ¶ 69. As discussed below, none of these cited references support this claim, and accordingly, a POSITA would not have concluded that reducing the resolution does not affect head detection and head pose estimation accuracy.

50. Dr. Doermann cites Ex. 1013 to indicate face detection can be a “resource-consuming” process. Ex. 1003, ¶ 69 (citing Ex. 1013, 000001). The notion of reducing the resolution of images is not addressed in Ex. 1013, let alone that such reduction would not affect the performance of head detection and head pose

estimation. Thus, Ex. 1013 does not support Petitioner's and Dr. Doermann's conclusion. Additionally, the "[v]iew-based [face detection] methods" quoted by Dr. Doermann as being computationally expensive and disclosed in Ex. 1013 are "**not appropriate**" for "**real-time applications**." Ex. 1013, 000001. As discussed above, the claimed audience measurement system of the '243 Patent analyzes the images in real-time in order to accurately measure the audience that is watching TV. *See* ¶ 13. Thus, a POSITA would not consider the resource-consuming view-based processes when modifying an audience measurement system, such as that of Lu-577.

51. Dr. Doermann also cites Ex. 1014 to support the opinion that "the processing resources and time can be reduced if the resolution of the image is reduced." Ex. 1003, ¶ 69 (citing Ex. 1014, 2:43-53). Specifically, Ex. 1014 states that "calculations are avoided of a complete highest resolution **integral image** for every acquired image in an image stream," thereby "reducing integral image calculations." *Id.* The "integral image" on which processing is applied is **derived from an acquired image**, as "each element of the integral image is **calculated** as the sum of intensities of all points above and to the left of the point in the image." Ex. 1014, 1:25-29, 2:33-35 (describing the method of the invention calculating a "corresponding integral image" for a portion of the image). Furthermore, the resource reduction is based on **reducing the resources required of integral image calculation**. Ex. 1014, 2:45-57. Creating this integral image would itself be a

resource-intensive process (under Dr. Doermann's definition), as, during the process, each pixel is created by summing intensities for every pixel above and left of the pixel being processed, as described above. Ex. 1014, 1:25-29. Furthermore, as the images captured in the modified system are not being processed in this way to create an integral image, any processing reductions disclosed in Ex. 1014 are irrelevant. This disclosed reduction is also irrelevant to any processing requirements of the images disclosed in the '243 Patent, as these images are unprocessed (excluding the resolution reduction by the pixel binner) prior to any head detection or head orientation determination. '243 Patent, 9:7-22 (describing the images analyzed to detect a head or orientation of the head are received "directly from the image capturer").

52. Ex. 1015, which is also cited by Dr. Doermann, is even less relevant to the proposed modification reducing resource consumption while not affecting head detection and head pose estimation accuracy. Ex. 1003, ¶ 69. Ex. 1015 describes reducing the resolution of the **saturation** of a "colour image" by "averaging the saturation." Ex. 1015, [0009]. The resolution of the saturation for each pixel is reduced. Ex. 1015, [0035] (describing each "pixel" has its own "averaged saturation value"), [0041] (describing an array of "reduced resolution pixels"). The method averages the saturation across pixels to make the colors in the image more uniform, such that a face may be more easily detected from background elements. Ex. 1015,

[0011-0015]. Dr. Doermann cites Ex. 1015, ¶ [0022], which states “by reducing the resolution **of the saturation of the image**, computational requirements are substantially reduced.” Ex. 1003, ¶ 69 (citing Ex. 1015, [0022]). However, the reduction in computational requirements is based on changing the resolution of the saturation component of each pixel (via saturation averaging), without changing the pixel resolution of an image. Additionally, Dr. Doermann ignores the increase in processing required to perform a conversion from RGB to an appropriate color space that calculates saturation, perform the reduction on the saturation channel, and then convert back to RGB for further processing. As Petitioner’s modification applies *Tian’s* reducing the pixel resolution of an image (rather than the saturation resolution), Ex. 1015 is irrelevant. Accordingly, none of Exs. 1013-1015 are relevant to the proposed modification, let alone provide evidence that reducing an image resolution does not affect the performance of head detection and head pose estimation.

53. Finally, Petitioner relies on *Tian’s* experimental results to support the claim that performing “face detection and pose estimation” on reduced-resolution images will not “substantially reduc[e] the accuracy of these tasks.” *Petition*, 10; *see also*, Ex. 1003, ¶ 69 (Dr. Doermann opining such tasks “do[] not substantially reduce the accuracy of these analysis tasks according to *Tian’s* teachings”). However,

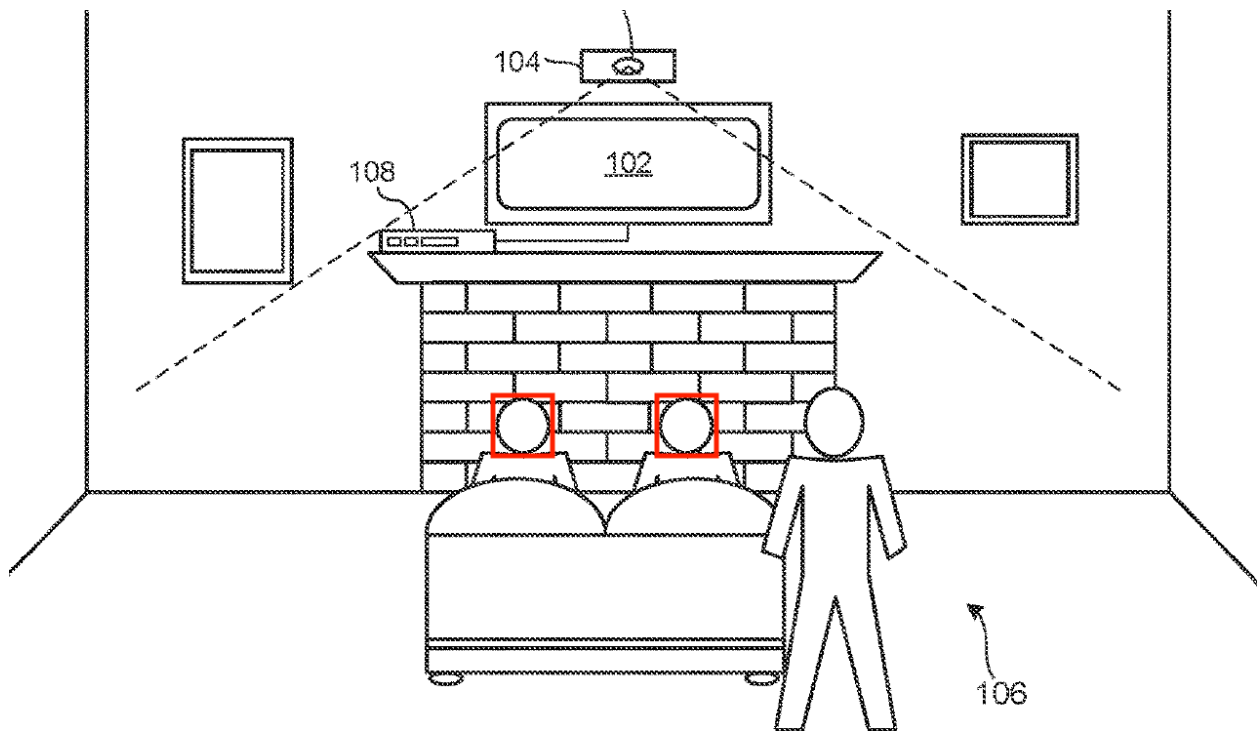
Tian's results **do not** indicate that there is no “substantial[]” accuracy reduction in the context of images captured in an audience measurement system.

54. Per *Tian*, the original images in the database are at a full resolution of 640x480 pixels. *Tian*, 000005. The “size of the head region” within these original images, as used by *Tian*, is 288x384. *Tian*, 000005, Table 3 (showing the “original” and highest-resolution image with a head region of 288x384). Thus, the head portion of the original images takes up about 36% of the entire image—or over a third of the image. Thus, *Tian* explores facial expression analysis—including face detection and head pose estimation—on images where a significant portion (over a third) of the image contains the face region.

55. Contrastingly, a face region within an image captured by an audience measurement system would take up a significantly smaller portion of the image. A camera of an audience measurement system captures an image of the entire media exposure environment (e.g., of the entire living room). '243 *Patent*, 1:58-62 (describing measurement systems capturing a series of images of a “media exposure environment (e.g., a television, room, a family room, a living room, a bar, a restaurant, etc.)”); *see also* claim 9 (reciting an audience measurement system collecting “a first image of the media exposure environment”). An exemplary media exposure environment is shown in FIG. 1 of the '243 *Patent* and reproduced below, with the field of view of the camera (shown with the dashed lines) extending to the

back of the room. The resolution of the captured image of the “environment” is, e.g., 1280x960. '243 *Patent*, 5:39-43.

56. Given the typical image parameters and the exemplary field of view shown in FIG. 1, a representative scaled image of the “media exposure environment” is shown below:



'243 *Patent*, FIG. 1, 5:9-13 (red boxes added). With the image scaled to 1280 pixels wide, the head region is ~60x60 pixels. For an image with resolution of ~1280x960 pixels, the head only takes up about 0.3% of the entire image. Comparing this head region of 60x60 pixels with *Tian's* accuracy results indicates to a POSITA any reduction of the image resolution would sacrifice accuracy of the head pose

estimation operation, as the head region is already below the minimum viable size (i.e., 72x96), even before the resolution is reduced. *Tian*, Table 3.

57. I note Figure 1 is applied in the '243 Patent for illustration purposes, and thus may not be to scale for an average TV-watching environment. For this reason, I searched for a real-world example of an image captured of an audience in a media exposure environment. The known average recommended viewing distance is ~10 feet away from a TV. *Is Your TV Too Big for Your Room?*, Ann Porter, October 3, 2007 (Ex. 2036) at p. 3. Given the average TV-watching distance, an exemplary image captured of a theoretical media environment by an audience measurement system is shown below:



The exemplary image is reproduced from this [website](#) (Ex. 2037). The orange dashed boxes annotate the adult “viewer’s” head/face region within the image, which is ~75x100 pixels. The blue boxes annotate the child viewer’s head/face region, which is ~65x80 pixels. With the original image being of size 1200x960, the face/head region takes up only about .65% of the original image for the adults, and .45% of the original image for the children. This is consistent with the results obtained for the media exposure environment depicted in FIG. 1 of the ’243 Patent.

58. The above image confirms the analysis of Figure 1 is representative of an actual media exposure environment. Comparing the adult face region of 75x100

pixels with *Tian*'s accuracy results indicates any **reduction** of the image resolution would sacrifice accuracy of the head pose estimation operation. Indeed, the full face region resolution approaches *Tian*'s minimum viable size (i.e., 72x96), even before the resolution is reduced. Namely, *Tian* found resolutions under 72x96 to affect the accuracy of head pose estimation. *Tian*, Table 3 (showing results of "97.8%" accuracy for the next resolution below 72x96). Additionally, for processes related to extracting features of the face, resolutions under 72x96 were found to decrease accuracy. *Tian*, 000005 (describing being able to extract features "well" for resolutions of 72x96 and higher). Seeing as the children's face/head region starts at 65x80 pixels, a POSITA would understand any reduction of the image resolution would sacrifice accuracy of the facial analysis operations (e.g., head pose estimation), because the original face region resolution is already below the minimum viable size (i.e., 72x96). I note that the claims of the '243 Patent do not include any particular resolution requirements; however, I performed this analysis to understand whether *Tian*'s conclusion would be applicable to images captured of a typical media exposure environment. I conclude that *Tian*'s conclusions would **not** be applicable in the context of images captured of a media exposure environment.

59. Thus, the starting resolution of the head region in an exemplary image captured by an audience measurement system is considerably lower than the resolution of *Tian*'s starting image. Furthermore, *Tian* shows degraded accuracy

occurs when face regions are at the original resolution size of $\sim 75 \times 100$ and reduced past that size. For these reasons, a POSITA **would not** find *Tian* to support Petitioner's purported statement that resolution reduction does not affect the accuracy of processes such as face detection and head pose estimation especially when applied to audience measurement systems. Rather, a POSITA would understand that *Tian's* results do not apply to performance accuracies of processing of images captured in an audience measurement system, as the starting resolution is already much lower than the starting full resolution explored in *Tian*.

b) The Proposed Modification to Perform Head Orientation Determination Is Flawed

60. Petitioner motivates modifying *Lu-577/Lu-000/Lert* to perform head orientation determination because doing so would “determine whether an image is suitable for face recognition” and thereby “improve[] the performance of face recognition.” *Petition*, 10. Petitioner further argues that determining head orientation would enable the system to determine “whether a person is watching the television.” *Id.* As discussed below, these purported benefits are not applicable in the context of audience measurement.

i. Petitioner's Modification Would Reduce Face Recognition Performance By Reducing Face Recognition Opportunities

61. Petitioner's combination of modifying *Lu-577/Lu-000/Lert* (and, specifically, modifying *Lu-000's* face detector) with *Tian's* teaching of determining

a head orientation is alleged to provide the benefit of identifying “suitable” candidate images before performing facial recognition to increase the likelihood of successful face recognition of an image. *Petition*, 10. However, such a benefit focuses on the chance of successful recognition of a particular image, **not** on the chance of successfully recognizing an audience member over all images. Consider an audience measurement system that captures 10 consecutive frames of video of an audience. Even with each of the 10 frames only having a 25% chance of success (e.g., given various head movements between frames that may obscure certain features), performing facial recognition on all 10 frames would still be more likely to identify an audience member than only a single face recognition operation on a single frame that has a 50% chance of success. Furthermore, a system that attempts to recognize faces for all frames will, by definition, also attempt to process the frames with the highest recognition chance. Thus, because having more images to perform facial recognition on increases the chances of recognizing an audience member, limiting the images by only identifying certain “candidate” images would make the recognition process worse, not better, as explained below.

62. Audience measurement systems were known to take video (i.e., a “series of images”) of a media exposure environment, and use these multiple image frames to “identify people” in the audience. *’243 Patent*, 1:56-64, 1:67-2:3. Thus, a POSITA would understand that many images of a potential audience member (e.g.,

via video) are captured by an audience measurement system. In such an audience measurement context, performing face recognition on **all** candidate images of the media exposure environment where the audience member is present maximizes the chance of successfully recognizing an audience member and accurately determining the media they are exposed to, as discussed below. Indeed, a POSITA would have understood various head poses, and not just those deemed “suitable” by Petitioner, would result in successful facial recognition. *Dec.*, ¶ 75. Thus, limiting the candidate images would decrease the chance of successful facial recognition.

63. Performing facial recognition on all candidate images to identify audience members was known, as *Lu-000* implements such an approach. Namely, *Lu-000* stores “face signatures” analyzed from multiple face images captured in a “series of video images” as reference signatures for each audience member. *Lu-000*, 7:25-38. Multiple reference signatures are stored for each audience member to account for various head poses. *Lu-000*, 8:36-40. Thus, when a captured image is compared to the reference signatures to “identify a match,” the system can recognize audience members regardless of the head orientation in the captured frame. *Lu-000*, 7:52-55.

64. Contrastingly, Petitioner’s modification would not use all captured images of the audience, but rather only certain images deemed “suitable.” Failing to attempt facial recognition on all images would decrease the accuracy of the audience

measurement data. For example, say a sequence of images are captured at 1 min into a TV program, 1 min 10 seconds into the TV program, 1 min 20 seconds into the TV program, etc. Treating all of those images as candidate images, the system can determine, with granularity, the time at which a recognized audience member was exposed to a certain media.

65. Compare this to only picking one image, such as the best image, for facial recognition. If only one image is used to recognize the audience member out of a sequence of images, the time intervals between candidate images increase and thus a user may be identified at, e.g., a 10 minute mark (when a best image was captured) and 20 minute mark (when another best image was captured). Between those times, the user may have gotten up from the exposure environment, but that would not be reflected in the audience measurements. Thus, a system that only attempts to recognize certain “suitable” frames, rather than all captured frames, would decrease the accuracy of the audience measurements. Thus, it is important to perform face recognition on all candidate images in the audience measurement context, in order to provide more precise measurements of times where an individual person is viewing media.

*ii. Petitioner’s Modification Would Not Allow
Determining When a Person Is Watching Television*

66. In addition to Petitioner’s motivation of improving recognition “accuracy” by determining “suitable” candidate images, Petitioner provides a

secondary motivation to perform head orientation determination in order to “determine whether a person is watching the television.” *Petition*, 10. To support the motivation, Dr. Doermann cites *Weinblatt’s* discussion of gaze tracking (i.e., tracking what the person is looking at) in determining if the person is watching TV. *Dec.*, ¶ 89. However, such gaze tracking is not implemented in *Tian*. *Weinblatt* discloses tracking the gaze by using (1) measures of the angle of a face, (2) determining if the eyes are opened or closed, or (3) determining if the face is “looking at” the TV. *Weinblatt*, [0039]. Contrastingly, *Tian* does not measure or determine **any** of these criteria required to determine “if the person is watching the broadcast program.” *Id.*

67. Rather, *Tian* applies head pose estimation to determine whether a head pose in an image is suitable for facial expression analysis. *Tian*, 000002-000003. As discussed previously, *Tian’s* head pose estimation via classification into three broad categories does not apply any of *Weinblatt’s* techniques for gaze tracking. *Dec.*, ¶ 89-94. Thus, a POSITA would understand *Tian’s* head pose estimation would not determine whether a person is watching TV.

68. Additionally, only recognizing audience members who are actively looking at the TV would **not** be beneficial in the context of audience measurement. This is the case at least because applying this narrow requirement would remove opportunities to recognize people who are still taking in the media content (i.e.,

audience members who are listening to it) but who are not looking at the time the image is taken. For example, *Lu-577*'s audience measurement system was developed to provide "changes in methods of measuring the reception of television and **radio programming**." *Lu-577*, [0007]. *Lu-577* indicates it is important not only to measure visual media viewed by an audience in an audience measurement system, but also auditory media heard by an audience. Thus, identifying audience members who are listening to the television broadcast, in addition to users who are otherwise watching the broadcast, is an important feature in an audience measurement system, which Petitioner's proposed modification would prevent.

69. A POSITA would appreciate that people standing/sitting in the media environment (e.g., living room) would still be "exposed" to the media on the TV, even if they are not directly viewing the media (e.g., not looking at the television). '243 *Patent*, 5:35-39 (describing the audience measurement system capturing images of the audience "being exposed" to media presented). Furthermore, media content was known to include "audio" and "radio," in addition to visual media, indicating to a POSITA that limiting an audience measurement system to identify only people watching the TV would make any measurements of media exposure inaccurate. '243 *Patent*, 1:43-45. Under Dr. Doermann's logic that an audience measurement system only need identify people who are looking at the TV, advertisers could reach their maximum audience by televising silent ads, given that

any person not actively looking at the television is not exposed to their advertisement. Rather, a POSITA would find it desirable to include all people in a room as the audience and thereby collect audience measurements for all people who can be recognized, even those not actively looking at the television.

c) Petitioner's Proposed Modification to Rely on Head Tracking Would Not Reduce the Net Processing Resources Required and Would Decrease the Performance of Face Recognition

70. As discussed above, Petitioner modifies *Lu-577/Lu-000/Lert* to include *Tian's* head tracking in a video sequence for teaching the two-frame process in dependent claims 5, 6, 12, 14, and 19. *See* ¶¶ 25-26. Petitioner motivates such a modification by stating doing so would “save[] processing resources and time by reducing the times face detection needs to be applied.” *Petition*, 9; Ex. 1003, ¶¶ 63-64. However, adding such head tracking into *Lu-577/Lu-000/Lert* (as otherwise modified) would require additional processing resources and incur other drawbacks:

- 1) Increased processing costs associated with performing head tracking;
- 2) No advantage in replacing head detection, as head detection on every frame would still be required to account for new audience members;
and
- 3) Decrease in facial recognition performance when using a head location from a first image to perform face recognition in the second image.

These drawbacks are discussed in further detail below.

71. Regarding drawback (1), Petitioner's motivation is that applying head/face tracking would decrease the times "face detection" needs to be applied. Even if face detection processes can be replaced by head/face tracking (which a POSITA would not agree with, as discussed further below), additional resources would be required to perform face tracking. For example, tracking requires detecting motion of an object within a sequence of image frames to track the position of the object (e.g., head/face). *Face Detection and Tracking Techniques: A Review*, Intl. Journal of Advanced Research in Computer Science, Vol. 2, No. 5, October 2011 (Ex. 2038) at p. 24. Techniques for determining movement may recognize movement of certain features or recognize movement of each pixel. *Id.* Regardless, to detect the movement, velocity vectors for points within the images are calculated to create a motion field and the vectors are used to determine the movement of the head/face. *Id.* at p. 25. These processes involved in tracking are not required in face detection. *See, generally, Face Detection and Tracking Techniques: A Review*, Intl. Journal of Advanced Research in Computer Science, Vol. 2, No. 5, October 2011 (Ex. 2038) (describing face detection techniques separate from face/head tracking techniques). Thus, a POSITA would understand head/face tracking requires additional resources in order to calculate the motion to accurately find the head in each image of a sequence.

72. Regarding drawback (2), no advantage for decreasing the amount of face detection applied exists, because face detection would still need to be applied to every image to detect any new audience members who entered the TV room. Indeed, even Petitioner's own reference applied for supporting the proposed modification teaches it is **"desirable" to "scan" each image "with a face detector" to "locate new faces entering the field of view of the camera,"** even when applying face/head tracking. *Petition*, 9 (citing Ex. 1010, [0003]; Ex. 1010, [0003]; *see also*, Ex. 1003, ¶¶ 63-64. Thus, a POSITA would have found Petitioner's proposed modification to **not provide savings in processing resources** by decreasing face detection instances, because head/face detection would still need to be performed in each new frame to locate new faces. The fact that resources are not reduced and increased resources are required to perform the head/face tracking suggests a POSITA would find the drawbacks to outweigh any benefit of the proposed modification. Thus, a POSITA would not be motivated to make the proposed modification.

73. Regarding drawback (3), a POSITA would understand performing facial recognition in a region of an image where the head was **in a prior frame** decreases the accuracy of facial recognition performance when compared to applying recognition in a region where the head **is located in the current frame**. For example, a POSITA would appreciate a person could move their head between

image captures in a sequence, and thereby render the previous location useless in recognizing the face in the current image. Indeed, *Tian* discloses the accuracy of facial analysis depends on whether the head is accurately tracked (i.e., “well-aligned”) in an image. *See* ¶ 30. Applying an up-to-date head location from the current image would ensure optimum head alignment over applying an out-of-date head location from a previous image. Thus, the accuracy of facial analysis improves by starting with an updated head location, rather than a location from a previous image.

74. To support the proposed modification using out-of-date head location to perform facial recognition in the current image, Dr. Doermann states the approach is “applicable when a sequence of images is taken for a person by a video camera” because “a person’s head or face is unlikely to move significantly.” Ex. 1003, ¶ 64. Dr. Doermann concludes “the region corresponding to the person’s head or face likely remains the same across the sequence of images.” *Id.* However, Dr. Doermann ignores that his relied-upon background references apply head tracking to locate the head accurately “from frame to frame.” Ex. 1003, ¶ 64 (citing Ex. 1010, [0003]). Namely, Ex. 1010 discloses “a localized region” around the located face “is scanned by a face detector **in the next frame**,” indicating to a POSITA the head/face must be relocated in a “next frame,” rather than just applying the previous location. Ex.

1010, [0003]. Thus, the cited reference does not support Dr. Doermann's conclusion that a previous location would be used to recognize a face when analyzing video.

75. Further, *Tian*, which Petitioner relies on for their two-frame process mapping, discloses that "large head motion" may occur between frames and, accordingly, "head tracking" is applied. *Tian*, 000001. This contradicts Dr. Doermann's unsupported statement that a person's head is "unlikely to move significantly" in "a sequence of images." Ex. 1003, ¶ 64. *Tian* applying head tracking to account for large head motion indicates to a POSITA that the head location may change enough to be significant, and thus head tracking must be applied to update the head location. *Tian* makes such a statement, even with the tested image set representing an optimum scenario for re-using a head position: a scenario where the images are of people standing still and looking at the camera to provide a facial expression. *Tian*, 000002-000003 (showing exemplary images where the head location is fairly static). This is because facial expressions do not require much movement of the head. Contrastingly, consider an audience measurement system where multiple audience members are sitting on a couch and may have "large head movement" occur between frames (e.g., audience members interacting with one another while being exposed to media). In this real-world scenario, the head location would need to be updated before performing face detection to accurately find the face.

76. *Tian* also states that even when the face image continues to be in the frontal/near frontal view in an image sequence, the location of facial features is tracked by “searching for the dark pixels around their positions in the last frame.” *Tian*, 000003. This indicates that even when little head movement occurs within an image sequence, the face still needs to be tracked. Given that both *Tian* and Ex. 1010 contradict Dr. Doermann’s conclusion, and that no references cited by Petitioner nor Dr. Doermann suggest that acceptable (let alone un-degraded) facial recognition can be performed using an un-updated head location from the previous frame, a POSITA would not understand that performing facial recognition in a region of an image where the head was previously located could be performed without degrading performance.

III. OPINIONS ON CHALLENGED GROUND 3

A. Opinions Regarding Modifying *Nielsen-372/Steinberg* with *Tian*’s Teachings

77. For Ground 3, Petitioner proposes further modifying *Nielsen-372* in view of *Steinberg* (introduced in now-moot Ground 2) to include *Tian*’s teachings to render the Challenged Claims obvious for Ground 3. As set forth below, for the same reasons applied to Ground 1, a POSITA would not have found it obvious or been motivated to modify *Nielsen-372/Steinberg* with *Tian*’s purported teaching of reducing a resolution of a first image or with *Tian*’s head/face tracking.

1. *Petitioner's Proposed Modification to Process Reduced-Resolution Images Would Reduce the Net Processing Resources Required*

78. Petitioner applies the same motivation for modifying *Nielsen-372/Steinberg* to include *Tian's* performing head pose estimation and face detection on reduced-resolution images as applied for Ground 1. *Compare Petition, 10 with Petition, 72.* Namely, Petitioner argues that such a modification would decrease resource consumption and time. *Id.* The analysis of such a modification is identical between the two Grounds. *Compare Petition, 72-73 with Petition, 10-11; compare Ex. 1003, ¶¶ 68-70 with Ex. 1003, ¶¶ 286-288.* Thus, similarly, Petitioner ignores the additional processing resources and time required to reduce the image resolutions and how this would outweigh the proposed benefits. For the reasons discussed regarding Ground 1 above in § II.C.1.a, a POSITA would not have found the proposed modification to provide the benefit of reducing processing resources and time, given the additional resources required to reduce the resolution and thus would not have been motivated to modify *Nielsen-372/Steinberg* with *Tian's* resolution reduction to teach claims 4, 11, and 18 (and via dependency, the other challenged claims).

2. *Petitioner's Proposed Modification to Rely on Head Tracking Would Not Reduce the Net Processing Resources Required and Would Decrease the Performance of Face Recognition*

79. Petitioner applies the same motivation for modifying *Nielsen-372/Steinberg* to include *Tian's* head tracking as applied for Ground 1. *Compare Petition, 9 with Petition, 72.* Namely, Petitioner argues that such a modification would decrease processing resources consumed and time. *Id.* The analysis of such a modification is identical between the two Grounds. *Id.* Thus, similarly, Petitioner ignores the additional processing resources required to perform head tracking and the disadvantages of decreasing the accuracy of face recognition performance by using an outdated previous head location for recognizing in a current image. For the reasons discussed regarding Ground 1 above in § II.C.1.c, a POSITA would not have found it obvious or been motivated to modify *Nielsen-372/Steinberg* with *Tian's* heading tracking to teach claims 5, 6, 12, 14, and 19 (and via dependency, claims 8, 13, and 20), given the drawbacks associated with such a modification outweighing the proposed benefit.

IV. CONCLUSION

I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and 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 Section 1001 of Title 18 of the United States Code.

Date: 4/30/2024

By: 
Eli Saber, Ph.D.