

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

META PLATFORMS, INC.; INSTAGRAM, INC.;
WHATSAPP LLC; META PLATFORMS TECHNOLOGIES, LLC;
and GIPHY, INC.
Petitioners

v.

VL COLLECTIVE IP LLC
Patent Owner

U.S. PATENT NO. 7,436,980

Case IPR2023-00924

**DECLARATION OF CHANDRAJIT BAJAJ, Ph.D. IN SUPPORT OF
PETITION FOR *INTER PARTES* REVIEW OF
U.S. PATENT NO. 7,436,980**

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	QUALIFICATIONS AND PROFESSIONAL EXPERIENCE	2
III.	LEVEL OF ORDINARY SKILL IN THE ART	7
IV.	RELEVANT LEGAL STANDARDS	8
V.	SUMMARY OF OPINIONS	13
VI.	THE '980 PATENT	14
A.	Technology Summary	15
B.	Claims	17
C.	Specification	18
D.	Summary of the Prosecution History	20
VII.	CLAIM CONSTRUCTION	22
VIII.	OVERVIEW OF THE PRIOR ART	23
A.	Zhao	23
B.	Steffens	25
C.	Ozer	30
D.	TLP	32
IX.	DETAILED ANALYSIS	34
A.	Ground I: Claims 1-16 Are Rendered Obvious Over Zhao	34
B.	Ground II: Claims 1-16 Are Obvious Over Steffens In View Of Zhao	68
C.	Ground III: Claims 1-16 Are Obvious Over Ozer In View Of Zhao	90
D.	Ground IV: Claims 1-16 Are Obvious Over TLP	106

X.	SECONDARY CONSIDERATIONS	120
XI.	CONCLUSION.....	120

LIST OF EXHIBITS

Exhibit No.	Description
1001	U.S. Patent No. 7,436,980 ('980 Patent)
1002	File History of U.S. Patent No. 7,436,980
1003	Provisional Application No. 60/574,799
1004	Declaration of Dr. Chandrajit Bajaj
1005	Curriculum Vitae of Dr. Chandrajit Bajaj
1006	Liang Zhao, Dressed Human Modeling, Detection, and Parts Localization, The Robotics Institute, Carnegie Mellon University (2001) ("Zhao")
1007	U.S. Patent No. 6,301,370 ("Steffens")
1008	U.S. Patent No. 7,200,266 ("Ozer")
1009	Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black & Michael Isard, <i>Tracking Loose-limbed People</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004) ("TLP")
1010	Program Documents for the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004), http://tab.computer.org/pamitc/archive/cvpr2004/index.html
1011	Adam Baumberg & David Hogg, <i>Generating Spatiotemporal Models from Examples</i> , School of Computer Studies, University of Leeds, http://www.bmva.org/bmvc/1995/bmvc-95-040.pdf ("Baumberg")
1012	Erik B. Sudderth, Alexander T. Ihler, William T. Freeman & Alan S. Willsky, <i>Non-parametric Belief Propagation</i> , IEEE Conference on Computer Vision and Pattern Recognition, (June 2003), https://www.ics.uci.edu/~ihler/papers/tr2551.pdf
1013	Declaration of Jessica Benner regarding "Zhao," Carnegie Mellon University (April 4, 2023)

Exhibit No.	Description
1014	Declaration of Carl Megeski regarding “Zhao,” ProQuest (March 2, 2023)
1015	Leonid Sigal, Michael Isard, Benjamin H. Sigelman & Michael J. Black, <i>Attractive people: Assembling loose-limbed models using non-parametric belief propagation</i> , Advances in Neural Information Processing Systems 16, NIPS 1539–1546 (MIT Press 2003)
1016	Hedvig Sidenbladh, Michael J. Black & David Fleet, <i>Stochastic tracking of 3D human figures using 2D image motion</i> , European Conference on Computer Vision, 702–718, LNCS 1843 (Springer-Verlag 2000)
1017	Noel A. C. Cressie, <i>Statistics for Spatial Data</i> , Chapters 1, 9, John Wiley and Sons (1993)
1018	Yaakov Bar-Shalom, X. Rong Li & Thiagalingam Kirubarajan, <i>Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software</i> , Chapters 1–2, 11, John Wiley and Sons, (2001)
1019	M. Isard, <i>PAMPAS: Real-valued graphical models for computer vision</i> , CVPR, vol. 1, at 613–620 (2003)
1020	D. L. Alspach & H. W. Sorenson. <i>Nonlinear Bayesian estimation using Gaussian sum approximations</i> , IEEE Trans. AC, 17(4): at 439–448, (August 1972)
1021	James M. Coughlan & Sabino J. Ferreira, <i>Finding deformable shapes using loopy belief propagation</i> , ECCV 453–468 (2002)
1022	F. De la Torre & M. J. Black., <i>Robust parameterized component analysis: Theory and applications to 2D facial modeling</i> , ECCV 653–69 (2002)
1023	S. L. Lauritzen, <i>Graphical Models</i> , Chapters 1, 3, Oxford University Press, Oxford (1996)

Exhibit No.	Description
1024	A. Doucet, N. de Freitas, & N. Gordon, editors. <i>Sequential Monte Carlo Methods in Practice</i> , Chapters 5, 16, Springer-Verlag, New York (2001)
1025	Vladimir Pavlovic, James M. Rehg, Tat-Jen Cham & Kevin Murphy, <i>A dynamic Bayesian network approach to figure tracking using learned dynamic models</i> , Intl. Conf. on Computer Vision 94–101 (1999)
1026	Y. Wu, G. Hua & T. Yu, <i>Tracking articulated body by dynamic Markov network</i> , Intl. Conf. on Computer Vision 1094–1101 (2003)
1027	Christopher Bregler & Jitendra Malik, <i>Tracking people with twists and exponential maps</i> , CVPR, pp. 8–15 (1998)
1028	J. Deutscher, A. Davison & I. Reid, <i>Automatic partitioning of high dimensional search spaces associated with articulated body motion capture</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition 669–676 (2001)
1029	Cristian Sminchisescu & Bill Triggs, <i>Covariance scaled sampling for monocular 3D body tracking</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, at 447–454 (2001)
1030	John MacCormick & Michael Isard, <i>Partitioned sampling, articulated objects, and interface-quality hand tracking</i> , European Conference on Computer Vision (2), at 3–19 (2000)
1031	Yacov Hel-Or & Michael Werman, <i>Constraint Fusion for Recognition and Localization of Articulated Objects</i> , International Journal of Computer Vision 19(1), at 5–28 (1996)

Exhibit No.	Description
1032	Y. Weiss & W. T. Freeman, <i>Correctness of belief propagation in Gaussian graphical models of arbitrary topology</i> , Neural Comp., 13:2173–2200 (2001)
1033	M. Fischler & R. Elschlager, <i>The representation and matching of pictorial structures</i> , IEEE. Trans. Computers, 22(1):67–92 (1973)
1034	M. Burl, M. Weber & P. Perona, <i>A probabilistic approach to object recognition using local photometry and global geometry</i> , European Conference on Computer Vision 628–41 (1998)
1035	P. Felzenszwalb & D. Huttenlocher, <i>Efficient matching of pictorial structures</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, at 66–73 (2000)
1036	D. Ramanan & D. Forsyth, <i>Finding and tracking people from the bottom up</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. II, at 467–716 (2003)
1037	Andrew Blake & Michael Isard, <i>Active Contours</i> , Chapters 7, 8, Springer-Verlag (1998)
1038	D. L. Alspach, <i>Gaussian Sum Approximations in Nonlinear Filtering and Control</i> , Information Sciences 7,271–290 (1974)
1039	A. R. Tilley, <i>The Measure of Man and Woman: Human Factors in Design</i> H.D. Associates (1993)
1040	Paul Viola & Michael Jones, <i>Rapid Object Detection using a Boosted Cascade of Simple Features</i> , Conference on Computer Vision and Pattern Recognition (2001)
1041	Hedvig Sidenbladh & Michael J. Black, <i>Learning Statistics Learning the Statistics of People in Images and Video</i> ,

Declaration of Chandra Bajaj, Ph.D. for IPR of U.S. Patent No. 7,436,980

Exhibit No.	Description
	International Journal of Computer Vision 54(1/2/3), at 183–209 (2003)
1042	Jonathan S. Yedidia, William T. Freeman & Yair Weiss, <i>Generalized Belief Propagation</i> , Advances in Neural Info. Proc. Sys. 13, at 689–95 (2000)
1043	<i>VideoLabs, Inc. et al v. Meta Platforms, Inc. et al</i> , Case No. 1-22-cv-00680, Dkt. 15 (D. Del. Aug. 29, 2022)

Unless otherwise indicated, all emphasis added.

I. INTRODUCTION

1. I am making this declaration at the request of Petitioners Meta Platforms, Inc.; Instagram, Inc.; WhatsApp LLC; Meta Platforms Technologies, LLC; and GIPHY, Inc. (“Meta” or “Petitioners”) in the *Inter Partes* Review of U.S. Patent No. 7,436,980 (“the ’980 Patent”).

2. I am being compensated for my work in this matter at my standard hourly rate of \$725. I am also being reimbursed for reasonable and customary expenses associated with my work and testimony. My compensation is not contingent on the outcome of this matter or the substance of my testimony.

3. I have been asked to provide my opinions regarding whether claims 1-16 (“the Challenged Claims”) of the ’980 Patent would have been obvious to a person having ordinary skill in the art (“POSITA”) at the time of the invention, in view of the prior art. It is my opinion that the Challenged Claims would have been obvious to a POSITA.

4. In the preparation of this declaration and forming my opinions, I have considered:

- the documents identified in the List of Exhibits table above;
- the relevant legal standards, including the standard for obviousness;
- any additional authoritative documents as cited in the body of this declaration; and

- my own knowledge and experience based upon my work in the areas of computer science, applied mathematics, image processing, computer vision and visualization as described below.

II. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE

5. My complete qualifications and professional experience are described in my *Curriculum Vitae*, a copy of which can be found in the attached Exhibit. EX1005. The following is a brief summary of my relevant qualifications and professional experience.

6. I am currently employed as a Professor of Computer Science at the University of Texas at Austin (“UT Austin”). I currently hold the Computational Applied Mathematics endowed Chair in Visualization. I am also the Director of the Computational Visualization Center at UT Austin, which has been funded by the National Institutes of Health, the National Science Foundation, the Department of Energy, and the Department of Defense. The center personnel include fifteen researchers, scientists, post-graduate students, and staff.

7. I have a Bachelor of Technology degree in Electrical Engineering, which I obtained from the Indian Institute of Technology in Delhi (IITD) in 1980. I also have a Master of Science degree and a Doctorate in Computer Science from Cornell University in 1983 and 1984 respectively.

8. Prior to my employment at the University of Texas, I was an assistant professor, then associate professor, and finally professor of Computer Sciences at Purdue University (Purdue) from 1984 until I resigned in 1997 and transferred to UT Austin. During this time, I was also the Director of Image Analysis and Visualization Center at Purdue University. I was a visiting associate professor of Computer Science at Cornell University from 1990–1991. I have also been invited for collaborative visits by several academic institutions and have presented numerous keynote presentations worldwide. I have been an editorial member of the SIAM Journal on Imaging Sciences, and the ACM Transactions on Graphics, and continue my editorial role for ACM Computing Surveys and the International Journal of Computational Geometry and Applications.

9. I have spent the better part of my career, both at Purdue and UT Austin, researching, designing, teaching, and using computer systems to model, simulate, and visualize natural and synthetic objects combining computational image and geometric processing. I am knowledgeable about and have much experience in both hardware and software, including algorithms, used for capturing, analyzing and displaying interactive imagery.

10. In the 1970s, while majoring in Electrical Engineering at Indian Institute of Technology with a minor in Computer Sciences, I was intimately involved in the design and fabrication of microprocessor-controlled circuits as well

Declaration of Chandra Bajaj, Ph.D. for IPR of U.S. Patent No. 7,436,980

as the development of microprocessor controller software. In the 1980s, while at Cornell University, my past experiences led to research in computational geometry, image processing, and optimization. In the early 1990s, I created 3D interactive and collaborative multimedia software environments which were fully navigable for multi-person computer gaming and simulation (e.g., “The GATI Client-Server Animation Toolkit” and “The VAIDAK Medical Image Model Reconstruction Toolkit”). In 1994, I co-authored a technical paper entitled “Shastra: Multimedia Collaborative Design Environment.” The need for increasing computer graphics display realism without sacrificing interactivity led me also to explore image processing and geometric modeling techniques such as segmentation and surface reconstruction from imaging scans, point clouds, and texture mapping with data compression, such as described in some exemplary publications “Arbitrary Topology Shape Reconstruction from Planar Cross Sections”, *Graphical Models and Image Processing*, 58(6):524-543 (1996); “Making 3D Textures Practical”, *Proceedings of Pacific Graphics*, 259-268 (1999); “3D RGB Image Compression for Interactive Applications,” *ACM Transactions on Graphics*, 20,1, 10-38 (2001); “Anisotropic Vector Diffusion in Image Smoothing”, *Proceeding of the 9th IEEE International Conference on Image Processing*, 828-831 (2002); “Image Segmentation Using Gradient Vector Diffusion and Region Merging”, *Proceedings of the 16th International Conference*

on Pattern Recognition, 941-944 (2002); “A Fast and Adaptive Algorithm for Image Contrast Enhancement”, *Proceedings of IEEE International Conference on Image Processing*, 1001-1004 (2004).

11. In the mid-2000s, I began to create spatially-realistic 3D graphical environments with a combination of different types of acquired and reconstructed imagery within which a user may explore, query, and learn. My publication titled “From Voxel Maps to Models” that appeared in an Oxford University Press book called “Imaging Life: Biological Systems from Atoms to Tissues”, is an example of my research in computational imaging. Over the course of my career, I have participated in the design and use of several computer systems including handhelds, laptops, graphics workstations to PC/Linux clusters, as well as very large memory supercomputers for capturing, modeling and displaying virtual and scientific phenomena. My experience with computer modeling and displaying computer graphics imagery encompasses many fields, such as interactive games, molecular, biomedical and industrial diagnostics, oil and gas exploration, geology, cosmology, and military industries.

12. During my time at UT Austin, I also designed and implemented scalable solutions for inverse problems in microscopy, spectroscopy, biomedical imaging; constructing spatially realistic and hierarchical 3D models; development of search/scoring engines for predicting energetically favorable multi-molecular

Declaration of Chandra Bajaj, Ph.D. for IPR of U.S. Patent No. 7,436,980

and cellular complexes; and statistical analysis and interrogative visualization of neuronal form-function. Additionally, I have courtesy appointments with, and supervise M.S. and Ph.D. students from, several UT departments, including, biomedical and electrical engineering, neurobiology, and mathematics. I currently serve on the editorial boards for the International Journal of Computational Geometry and Applications, and the ACM Computing Surveys. Much of my work involves issues relating to interactive image processing, 3D modeling, bioinformatics, computer graphics, and computational visualization. Examples of my publications, including peer-reviewed publications, are listed in my CV.

13. As set forth in my CV, I have authored approximately 170 peer-reviewed journal articles, 34 book chapters (which were also peer reviewed), and 157 peer-reviewed conference publications.

14. I have written and edited four books, on topics ranging from image processing, geometric modeling and visualization techniques, to algebraic geometry and its applications. I have given 215 invited speaker keynote presentations. I am a Fellow of the American Association for the Advancement of Science (AAAS), a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), a Fellow of the Society of Industrial and Applied Mathematics (SIAM), and also a Fellow of the Association of Computing Machinery (also known as ACM), which is the world's largest education and scientific computing society.

The ACM Fellow is ACM's most prestigious member grade and recognizes the top 1% of ACM members for their outstanding accomplishments in computing and information technology and/or outstanding service to ACM and the larger computing community.

III. LEVEL OF ORDINARY SKILL IN THE ART

15. I understand there are multiple factors relevant to determining the level of ordinary skill in the pertinent art, including (1) the levels of education and experience of persons working in the field at the time of the invention; (2) the sophistication of the technology; (3) the types of problems encountered in the field; and (4) the prior art solutions to those problems.

16. A person of ordinary skill in the art ("POSITA") in the field of the '980 Patent, as of its earliest possible priority date of May 27, 2004, would have been someone with a bachelor's degree in computer science, computer engineering, computer vision or visualization, or an equivalent field, and approximately two years of experience in software development for computer vision applications. Additional education might compensate for less experience, and vice-versa.

17. I determined the level of skill of a person of ordinary skill in the art by reviewing the '980 Patent and considering the type of problems encountered in the art, such as the purported need for "a system and method of detecting and

tracking an object, implementing component detection and performing inference over space and time.” *See* EX1001 at 1:22-24. Solutions to these problems involve algorithm design and software development. These topics are routinely covered in multiple undergraduate curriculum courses in computer science, computer vision, visualization, and computer engineering. Problem solving application expertise is gained through projects and courses and industrial experience in real-time computer vision, visualization, interactive computer graphics, image processing and pattern recognition, data analysis and visualization.

18. Based on my own education and experience, by the earliest possible priority date of May 27, 2004, I was at least a POSITA and I am familiar with the degree of knowledge that a POSITA would have possessed and how a POSITA would have interpreted the '980 Patent.

IV. RELEVANT LEGAL STANDARDS

19. I am not an attorney. In preparing and expressing my opinions and considering the subject matter of the '980 Patent, I am relying on certain basic legal principles that counsel have explained to me. These principles are discussed below.

20. I understand that a person of ordinary skill in the art (a “POSITA”) is a hypothetical person who is presumed to have known the relevant art at the time of the invention, and that the factors that may be considered in determining the

level of ordinary skill in the art may include: (a) type of problems encountered in the art; (b) prior art solutions to those problems; (c) rapidity with which innovations are made; (d) sophistication of the technology; and (e) educational level of active workers in the field. In a given case, every factor may not be present, and one or more factors may predominate.

21. I have been informed that patent claims are construed or interpreted from the perspective of a POSITA at the time of the claimed invention. I understand that the most important evidence to consider in construing the claims is the “intrinsic” evidence. This includes the claim language, the patent’s specification, and the prosecution history (including references cited in the prosecution history). I understand that the POSITA must read the claim terms in the context of the claim and in the context of the entire patent specification. I understand that the patentee may explicitly define a claim term in a way that differs from the plain and ordinary meaning in the art.

22. I further understand that a POSITA may also consider “extrinsic” evidence when interpreting the meaning of claim terms, which are documents and testimony beyond the patent to ensure that a claim is construed consistently with the understanding of those of skill in the art at the time of the claimed invention. I also understand that extrinsic evidence may not be relied on if it deviates from the meaning of the claim provided by the intrinsic evidence.

23. I further understand that inventors may draft their claims as a means or step for performing a specified function, but that they must also disclose in the specification the structure that performs the function, or that claim is indefinite. I also understand that construing a means-plus-function claim requires identifying the claimed function and determining what structure, if any, disclosed in the specification corresponds to the claimed function. I further understand that a term with a computer-implemented function must have a clearly linked structure in the form of an algorithm. I understand that if I cannot identify such an algorithm, the term lacks a clearly linked structure and is therefore indefinite. I further understand that a general-purpose computer cannot be the structure for a computer-implemented function because this amounts to purely functional claiming.

24. I understand that a claim is invalid if it is anticipated or obvious. Anticipation of a claim requires that every element of a claim be disclosed expressly or inherently in a single prior art reference, arranged in the prior art reference as arranged in the claim.

25. It is my understanding that a patent claim is invalid for obviousness if the claimed invention would have been obvious at the time the invention was made to a person having ordinary skill in the art (POSITA).

26. I understand that the following factors are considered in an obviousness analysis: (1) the scope and content of the prior art; (2) the differences

between the art and the claims at issue; (3) the level of ordinary skill in the art at the time the invention of the asserted patent was made; and (4) objective evidence of nonobviousness. In order to determine whether claim elements are found in the prior art, it is necessary to compare the properly construed claim language of the patent with the teachings of the prior art.

27. I understand that certain factors—often called “secondary considerations” may support or rebut an assertion of obviousness of a claim. I understand that such secondary considerations include, among other things, commercial success of the alleged invention, skepticism of those having ordinary skill in the art at the time of the alleged invention, unexpected results of the alleged invention, any long-felt but unsolved need in the art that was satisfied by the alleged invention, the failure of others to make the alleged invention, praise of the alleged invention by those having ordinary skill in the art, and copying of the alleged invention by others in the field. I further understand that there must be a nexus—*i.e.*, a connection—between any such secondary considerations and the alleged invention.

28. I further understand that a claim can be found obvious if it unites old elements with no change to their respective functions or alters prior art by mere substitution of one element for another known in the field, with that combination yielding predictable results. While it may be helpful to identify a reason for this

combination, common sense should guide, and there is no rigid requirement for a teaching, suggestion, or motivation to combine. For example, design incentives and other market forces can prompt variations of a product, either in the same field or different one. If a person having ordinary skill in the relevant art can implement a predictable variation, obviousness can bar patentability. Similarly, if a technique has been used to improve one device, and a person having ordinary skill in the art would recognize that the technique would improve similar devices in the same way, use of the technique is obvious. I further understand that a claim may be obvious if common sense directs one to combine multiple prior art references or add missing features to reproduce the alleged invention recited in the claims.

29. I also understand that any one or more of the following rationales may support a finding of obviousness:

- Combining prior art elements according to known methods to yield predictable results;
- Simple substitution of one known element for another to obtain predictable results;
- User of known technique to improve similar devices (methods, or products) in the same way;

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

30. I also understand that it is impermissible to use the claimed invention itself as a blueprint for piecing together elements in the art. In other words, it is impermissible to use hindsight reconstruction to pick and choose among disclosures in the prior art to reconstruct the claimed invention.

V. SUMMARY OF OPINIONS

31. I discuss and cite many documents through my declaration in support of my analysis, but my opinion regarding the obviousness of the Challenged Claims is fundamentally based on:

- **“Dressed Human Modeling, Detection, and Parts Localization” by Liang Zhao (“Zhao”) (EX1006)**, a dissertation submitted to The Robotics Institute of Carnegie Mellon University. Considering when Zhao was publicly available, I understand that it is prior art to the ’980 Patent.
- **U.S. Patent No. 6,301,370 (“Steffens”) (EX1007)**, filed on December 4, 1998 and granted on October 9, 2001. I understand that based on Steffens’ filing date, it is prior art to the ’980 Patent.
- **U.S. Patent No. 7,200,266 (“Ozer”) (EX1008)**, filed on August 27, 2003 and granted on April 3, 2007. I understand that, based on Ozer’s filing date, it is prior art to the ’980 Patent.
- **“Tracking Loose-limbed People” by Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard (“TLP”) (EX1009)**, a paper submitted for the IEEE Computer Society Conference on Computer Vision and Pattern Recognition in 2004. Considering when TLP was publicly available, I understand that it is prior art to the ’980 Patent.

32. In my opinion, Claims 1–16 of the ’980 Patent would have been obvious to a POSITA before May 27, 2004, as summarized in the following table:

Ground	Claims	Analysis
I	1–16	Obvious under §103 over Zhao
II	1–16	Obvious under §103 over Steffens in view of Zhao
III	1–16	Obvious under §103 over Ozer in view of Zhao
IV	1–16	Obvious under §103 over TLP

VI. THE ’980 PATENT

33. I understand that the ’980 Patent issued on October 14, 2008, from U.S. Application No. 11/135,210, which was filed on May 23, 2005. Further, the ’980 Patent claims priority to Provisional Application No. 60/574,799, which was filed on May 27, 2004.

A. Technology Summary

34. Before the '980 Patent's priority date, computer vision was a well-known field in image processing where computers were used to detect, recognize, or track objects in images. EX1001 at 1:10-17; EX1003 at 4 ("The problem of describing and recognizing categories of objects (e.g., faces, people, cars) has been an active field of computer vision for a considerable time."). For computer vision applications performing object detection, it was and remains "common to represent objects as collections of features with distinctive appearance, spatial extent, and position." EX1001 at 1:14-19; EX1003 at 4 ("A seemingly global agreement has been reached on the representation of object categories, with most researchers following the trend of representing objects as collections of features with distinctive appearance, spatial extent, and position.").

35. According to the intrinsic record of the '980 Patent, known methods for object detection included "the use of graphical models and belief propagation...." EX1003 at 5 ("We model objects using two-layer spatio-temporal graphical model, and use non-parametric belief propagation to infer the position and scale of the object and components. We should note that the use of graphical models and belief propagation is not novel for object detection."). Indeed, these methods, which also included the use of spatio-temporal models, were well known to a POSITA. *Id.*; *see also* EX1003 at 5 (other known techniques included

“temporal propagation”). For example, graphical models (*see* EX1023) have been well known to provide an efficient, powerful framework for modeling articulate spatio-temporal objects (e.g., humans, cars, etc.) using probabilistic relationships. *See* EX1032. In computer vision applications such as video human tracking, graphical models employ approximate inference algorithms such as Bayesian optimization (BO) and loopy belief propagation (BP) used in Alspach (*see* EX1020) and Isard (*see* EX1019) to infer the position and scale of the object and components. The BO and BP approximate inference methods produce good results for a wide range of articulate models as also shown in papers by Coughlan et. al. (*see* EX1021) and De la Torre (*see* EX1022). For temporal inference problems defined on Markov chains, and temporal graphical models, a variety of spatio-temporal object feature density representations have been explored over a moving window of video image frames. These probabilistic methods include Gaussian sum filters (*see* EX1020) that use local linearizations to deterministically propagate Gaussian mixture models in space and time through the video image sequence frames. In contrast, particle filters (*see* EX1024) use Monte Carlo methods to stochastically update a set of weighted point samples of feature density representations of probabilistic graphical models within the temporal window of video image sequences.

36. Additionally, to perform object detection, it was known to a POSITA to use probabilistic frameworks for determining the presence of an object in an image or sequence of images. *See* EX1003, 9 (describing the use of “image likelihood,” *i.e.*, “the likelihood of observing the [object in the] image”). For example, approaches to human detection and tracking exploit articulated spatio-temporal body models in either 2.5D (*see* EX1025), or 3D (*see* EX1016, EX1026-EX1029) where D stands for real dimensions, lead to a parametric state-space representation of spatio-temporal features of roughly 25–35 dimensions. The high dimensionality of the resulting feature space requires stochastic search algorithms to infer the position and scale of the objects that exploit the highly redundant dynamics of typical human motions (*see* EX1016), or use hierarchical sampling schemes to exploit the tree-structured nature of the spatio-temporal model. *See* EX1028, EX1030.

B. Claims

37. The '980 Patent has 16 claims, including 2 independent claims and 14 dependent claims. Claims 1-8 are substantially similar to Claims 9-16. Claim 1 closely matches Claim 9, except for minor differences to the preamble. Claims 2-8 mirror Claims 10-16, except that Claims 2-8 depend on Claim 1 whereas Claims 10-16 depend on Claim 9.

C. Specification

38. The '980 Patent claims a probabilistic framework for “component-based detection and tracking of objects in images and/or video” that “combines object detection with tracking.” EX1001 at 2:43-47. The method and steps of the '980 Patent is illustrated as a flowchart in Figure 5:

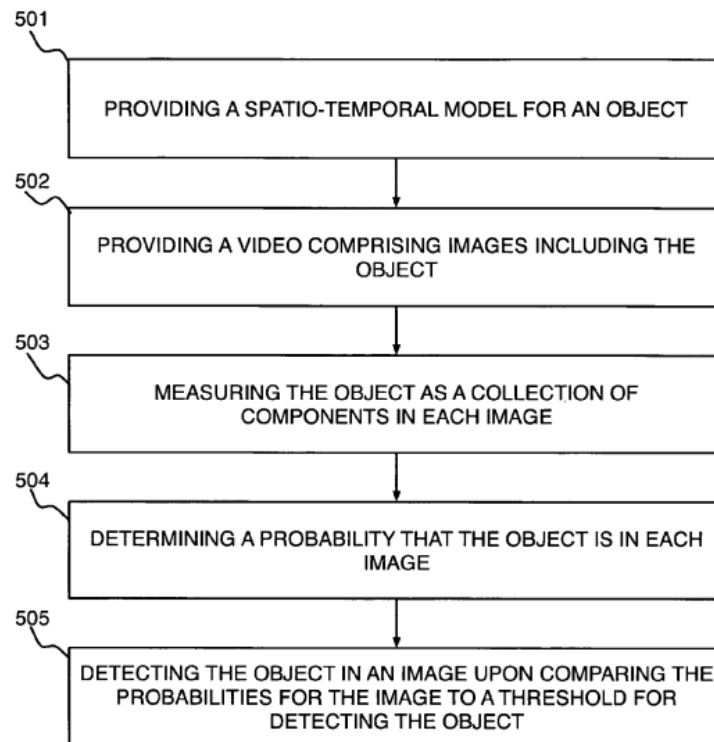


FIGURE 5

EX1001 at Fig. 5.

39. The '980 Patent begins with providing a “spatio-temporal model” for an object to be detected. EX1001 at Fig. 5; Claims 1, 9. Next, the method further comprises “providing a video comprising images including the object” and “measuring the object as a collection of components in each image.” *Id.* To detect

the object, the '980 Patent uses a probabilistic framework by “determining a probability that the object is in each image” and “detecting the object in an image upon comparing the probabilities for the image to a threshold for detecting the object.” *Id.*

40. With respect to the “spatio-temporal model,” the '980 Patent specification states “[t]he spatio-temporal model is a graphical model comprising nodes corresponding to each of the collection of components and to the object.” EX1001 at 1:40-42. For example, objects are represented using a graphical model, as illustrated in Figure 2A, which is annotated below:

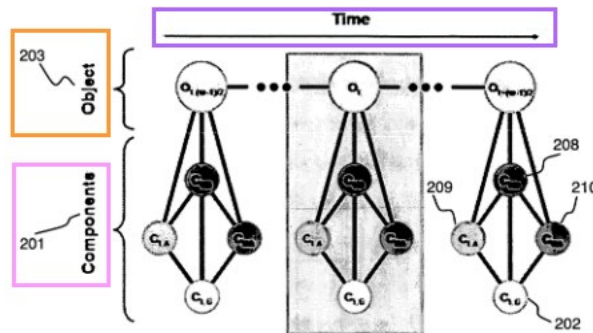
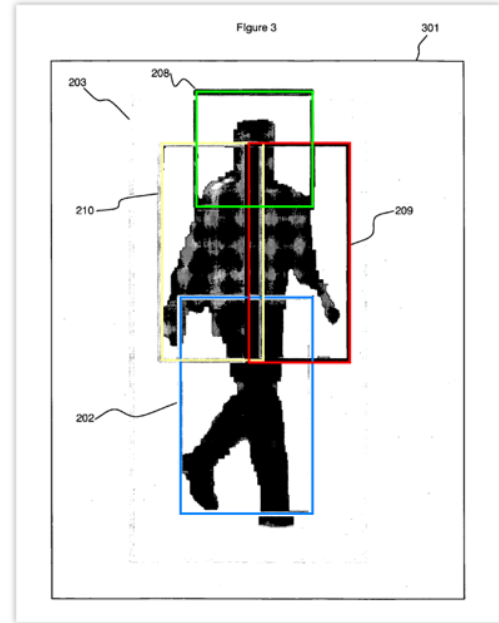
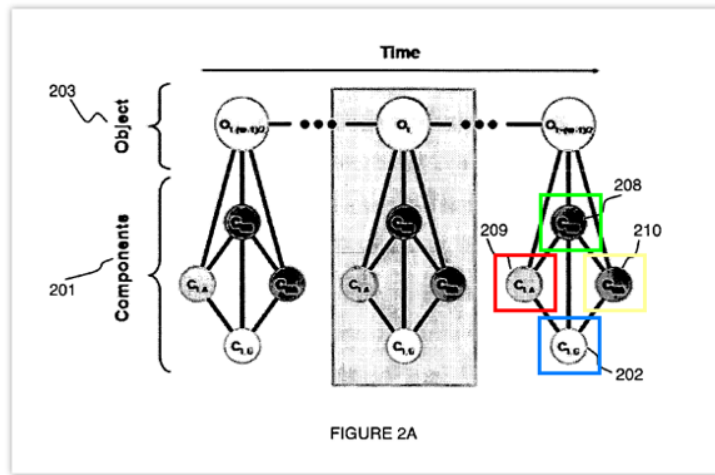


FIGURE 2A

EX1001, Fig. 2A.

41. As illustrated above, the “object” layer is represented by orange-boxed element 203. The object is comprised of parts in the “component” layer represented by pink-boxed element 201. To integrate temporal elements, the model is extended over time, as illustrated in purple. The model corresponds to an object in an image:



EX1001, Figs. 2A, 3.

42. As illustrated above, the model corresponds to an object in an image. For example, “for an image 301 of the pedestrian, they are: head (HD) 208 [green], left arm (LA) 209 [red], right arm (RA) 210 [yellow] and legs (LG) 202 [blue] (see FIG. 3).” EX1001, 4:31-33.

D. Summary of the Prosecution History

43. I understand that during prosecution of the '980 Patent, the examiner recognized the existence of numerous, “close prior art” references. EX1002 at 163. Effectively, the examiner did not identify any individual limitation as novel or nonobvious, but instead the examiner found “no suggestion or motivation to combine” the prior art references and the “features in combination” were purportedly not found in the record. EX1002 at 162-163.

44. First, the examiner acknowledged that the “close prior art, U.S. Patent No. 6,597,801 B1, Cham et al, teaches providing a kinematic model for an object to be detected (column 11, line 26), providing a video comprising a plurality of images including the object (figure 4B); measuring the object as a collection of components in each image (figure 10A); performing feature search operations for object features by comparing a total probability with a predetermined threshold (column 3, lines 64-65).” EX1002, 163. While the examiner did not dispute that Cham disclosed the probabilistic framework of the ’980 Patent, the examiner found that Cham did not describe a spatio-temporal model: “However, Cham does not teach the claimed ‘spatio-temporal model’, which is different from the kinematic model (spatial model).” *Id.*

45. Next, the examiner acknowledged that “another close prior art, U.S. Patent No. 7,200,266 B2, Ozer et al, teaches a method to detect objects in multiple video frames through object modeling and graph matching, where the output of the graph matching is the classified object parts, the movement of the object parts are a spatio-temporal sequence of feature vectors that consist of the direction of the object part movement.” EX1002 at 163. “Additionally, Ozer’s system checks direction of the movements of the object parts for a number of frames and calculates the probabilities of the activ[iti]es with the known activities by using Hidden Markov Models and choose the pattern with the highest probability as the

recognized activity in these frames.” *Id.* While the examiner did not dispute that Ozer taught a spatio-temporal model, the examiner found that it did not teach the probabilistic framework claimed in the ’980 Patent: “However, Ozer does not teach the claimed features of determining a probability that the object is in each image; and detecting the object in any image upon comparing the probabilities for each image to a threshold for detecting the object.” *Id.*

46. While each element of the ’980 Patent was found in the prior art, the examiner found that “there is no suggestion or motivation to combine the Cham and Ozer references...” EX1002 at 162-163. Thus, the “features in combination” were purportedly not found in the prior art record, so the claims were allowed. *Id.*

VII. Claim Construction

47. I understand that, in parallel litigation between the parties, it appears that Patent Owner may seek to construe the term “spatio-temporal model.” In Patent Owner’s opposition briefing to a motion to dismiss under Section 101, I understand that Patent Owner contended: “The ’980 patent’s spatio-temporal model is a graphical model with nodes that correspond to each of the collection of components and to the object.” *VideoLabs, Inc. et al v. Meta Platforms, Inc. et al*, Case No. 1-22-cv-00680, Dkt. 15 at 6 (D. Del. Aug. 29, 2022) (citing ’980 patent, 1:41–42). Currently, I present no opinion on claim construction, as I believe the

Challenged Claims would have been obvious even under Patent Owner's contended construction.

VIII. OVERVIEW OF THE PRIOR ART

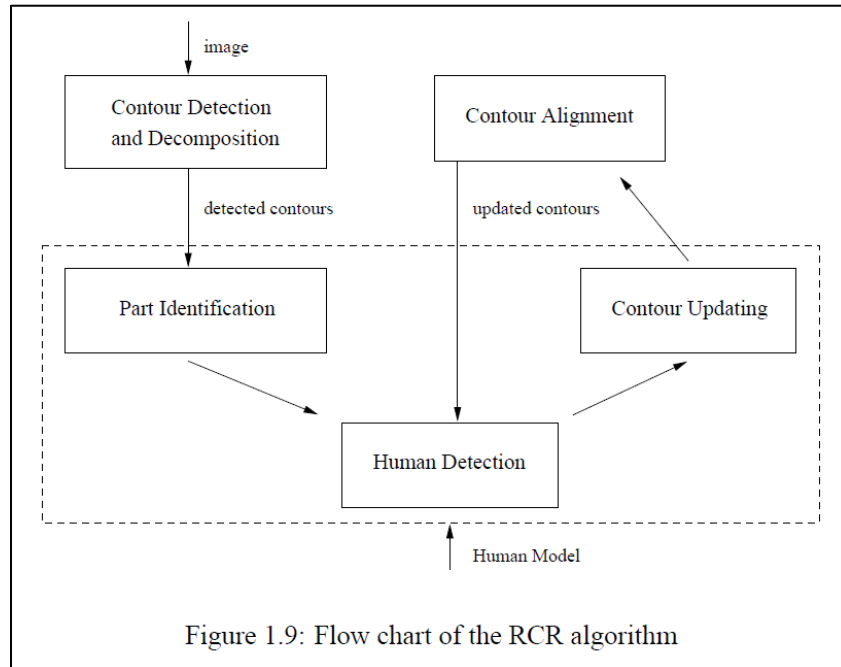
A. Zhao

48. Zhao is entitled "Dressed Human Modeling, Detection, and Parts Localization." Zhao is a dissertation submitted to The Robotics Institute of Carnegie Mellon University. *See* EX1006 at 1. The date listed on the face of Zhao is July 26, 2001. I understand that Zhao was publicly available no later than September 23, 2002, according to the declaration provided by Carnegie Mellon University. *See* EX1013 at 3 ("the Zhao Dissertation would have been accessible to the public at the Engineering & Science Library by September 23, 2002").

Additionally, I understand that Zhao was also stored in the ProQuest repository and made publicly available at least as early as June 25, 2002. *See* EX1014 at 3 ("Date first made available for sale (print or electronic)[:] 06/25/2002").

49. Zhao is a paper describing a computer-implemented method and system for object detection, including automatically detecting pedestrians, and locating their joints, in various shapes, sizes, postures, partial occlusion, and clothing. As part of its method, Zhao presents what it refers to as a "recursive context reasoning" ("RCR") algorithm to accurately locate the joints between body parts as part of "an integrated human shape modeling, detection, and body part

localization vision system.” EX1006 at 26, 97. The RCR algorithm is depicted in a flowchart as illustrated in Figure 1.9:



EX1006 at 26.

50. As shown above, Zhao begins with an image and performs “Contour Detection and Decomposition”—extracting contours of objects in image data and decomposing them into parts. EX1006 at 26, 70-71. Then, Zhao performs “Part Identification” and “Human Detection” where decomposed parts are then matched to an object model to calculate the probability of the object’s presence in the images. *Id.* If the calculated probability surpasses a defined threshold, the object is detected. *Id.*

51. Zhao also describes an object model, which is specifically a “human model.” This model is also a spatio-temporal model with spatial and temporal

elements. Zhao’s “model is designed to encode the shapes of the body parts and the context information — the size and *spatial relationships* between body parts.”

EX1006 at 69. Further, as Zhao recognizes, when detecting articulated objects such as humans, overlapping body parts can create ambiguity and those body parts be confused with one another, which may be viewed as a form of occlusion. Zhao’s spatio-temporal model addresses this issue as “[m]otion information obtained from previous frames can be used to predict the orientations of the limbs and to solve the ambiguity.” EX1006 at 91.

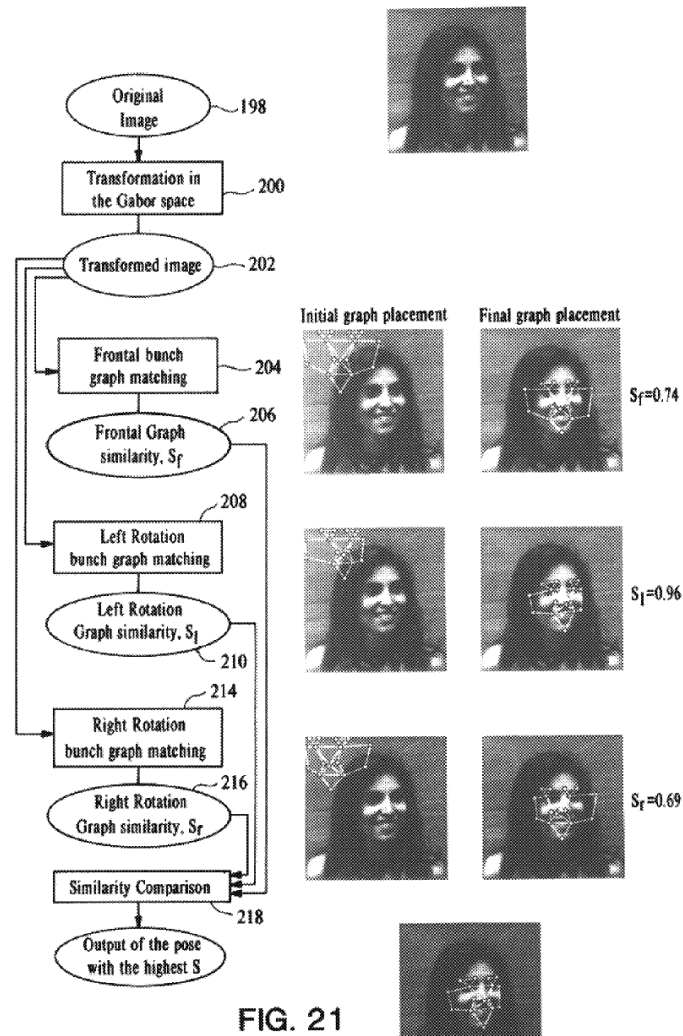
52. Zhao’s method for object detection is also a probabilistic framework. Zhao uses a probabilistic framework known as a Bayesian similarity measure, which combines local shape and global relationship constraints into a single equation to evaluate the degree of resemblance between a contour and the assembled human model. EX1006 at 25.

53. Finally, Zhao’s method uses the initial measurements to update the original contour and improve the accuracy of its object detection, thus forming a “recursive context reasoning” algorithm.

B. Steffens

54. Steffens is entitled “Face Recognition From Video Images” and is a US patent that was filed on December 4, 1998 and granted on October 9, 2001.

55. Steffens details a method for “detecting objects in video images, such as human faces, and tracking and identifying objects in real time.” EX1007 at 1:14-17. According to Steffens, “[t]he object identification and recognition process uses an image processing technique based on model graphs and bunch graphs that efficiently represent image features as jets.” *Id.* at 1:44-47. In an embodiment of Steffens, a process called elastic bunch graph matching is used to perform facial recognition, which is a form of object detection. This process is illustrated in Figure 21:



EX1007 at Fig. 21.

56. As illustrated above, Steffens begins its object detection and facial recognition process by transforming images in Gabor space, such as elements 198-202, and then performs “bunch graph matching” on the image, such as elements 204, 208, and 214, to perform “similarity comparisons,” such as in elements 206, 210, 216, and 218.

57. A bunch graph according to Steffens is illustrated in Figure 11, and is a data structure like an image graph, which “is used to describe the aspects of an object.” EX1007 at 8:56-57, 9:22-25.

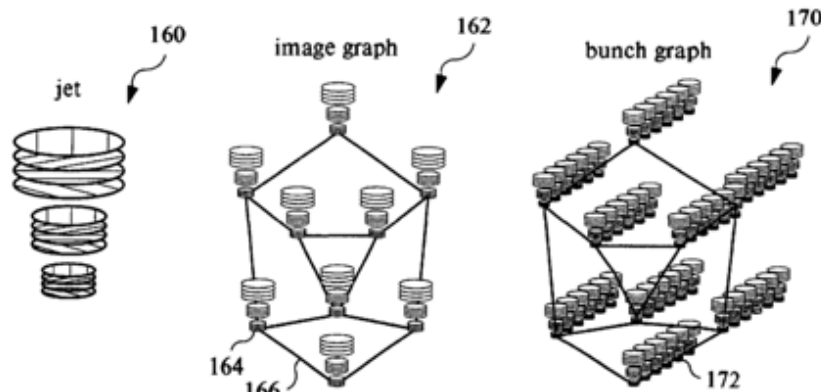


FIG. 11

EX1007 at Fig. 11.

58. As shown above, the nodes of the labeled graph are points on the object and are labeled with “jets,” and “[e]dges of the graph are labeled with distance vectors between the nodes.” EX1007 at 8:58-60. Each jet describes the local features of the area surrounding an image point. EX 1007 at 8:45-55. All Gabor-based wavelet components centered in a single image point and considered as a vector is called a jet. Each jet describes the local features of the area surrounding an image point. EX 1007 at 8:45-55. With respect to a bunch graph, as opposed to an image graph, “instead of attaching only a single jet to each node, a whole bunch of jets 172 (a bunch jet) are attached to each node.” EX1007 at 9:22-25. Facial images are used to derive jets, and to form a bunch graph, “a

collection of facial images (the bunch graph gallery) is marked with node locations at defined positions of the head,” which are called “landmarks.” EX1007 at 9:25-29.

59. Steffens explains, “[w]hen matching a bunch graph to an image, each jet extracted from the image is compared to all jets in the corresponding bunch attached to the bunch graph and the best-matching one is selected. This matching process is called elastic bunch graph matching.” EX1007 at 9:29-34. When “the object is in a sequence of images ... the step of detecting an object further includes tracking the object between image frames based on a trajectory associated with the object. Also, the step of locating the nodes includes tracking the nodes between image frames and reinitializing a tracked node if the node’s position deviates beyond a predetermined position constraint between image frames.” EX1007 at 2:6-13. Accordingly, Steffens claims as part of its process for recognizing objects, “tracking the object between image frames based on a trajectory associated with the object,” “tracking the nodes between image frames,” and “reinitializing a tracked node if the node’s position deviates beyond a predetermined position constraint between image frames.” EX1007 at Claims 14, 16-17.

C. Ozer

60. Ozer is entitled “Method And Apparatus For Automated Video Activity Analysis” and is a US patent that was filed on August 27, 2003 and granted on April 3, 2007.

61. As described in Ozer, it teaches a method to “detect, recognize, and analyze people or other objects” in video images. EX1008 at Abstract. Ozer’s method can be applied in the context of “security check points, public-places, parking lots, or in similar environments under surveillance to detect the presence of certain objects of interests (e.g., people), and to identify their activities for security and other purposes in real-time.” *Id.* As Ozer details, “[o]bject detection is achieved by matching the relational graphs of objects with the reference model. This method maps the attributes, interprets the match and checks the conditional rules in order to index the parts correctly.” EX1008 at 4:24-27. “For example, if the object of interest is a human body, the system will first detect different object parts, e.g. hands, head, arms, legs, torso and compare these part attributes with the human model attributes via graph matching.” EX1008 at 4:38-41.

62. When performing object detection, each extracted part of an image “corresponds to a node in the graphical representation of the object of interest. ... The combination of segments is controlled by the reference model and by the rule

generator.” EX1008 at 7:8-31. Accordingly, “[t]he output of the graph-matching algorithm is the classified object parts.” *Id.* at 7:34-35.

63. Further, Ozer teaches “Classifying Over Multiple Frames” where “[t]he movements of the object parts are described as a spatio-temporal sequence of feature vectors that consist of the direction of the object part movement. The system checks direction of the movements of the object parts for a number of frames and calculates the probabilities of the activities with the known activities by using Hidden Markov Models and chooses the pattern with the highest probability as the recognized activity in these frames (408). The activity of each part is then combined by a quadratic distance classifier to find the overall activity of the object of interest.” *Id.* at 7:33-45.

64. For context on “Hidden Markov Models,” in probability theory and related fields, a stochastic or random process is usually defined as a sequence of random variables where the index of the sequence have the interpretation of time. Stochastic processes are widely used as temporal models of noisy systems and phenomena that appear to vary in a random manner. A Markov chain or Markov process is a stochastic process describing a sequence of possible events in which the probability of each event depends *only on* the state attained in the *previous* event. It is named after the Russian mathematician Andrey Markov. A hidden Markov model (HMM) is a probabilistic Markov model in

which the system being modeled is assumed to be a Markov process—call it X with unobservable (“*hidden*”) states. As part of the definition, HMM requires that there be an observable process Y whose outcomes are “influenced” by the outcomes of X in a known way. Since X cannot be observed directly, the goal is to learn about X by observing Y. HMM has an additional requirement that the outcome of Y at time $t=t_0$ must be “influenced” exclusively by the outcome of X at $t=t_0$, and that the outcomes of X and Y at time $t < t_0$ must be conditionally independent of Y at $t=t_0$ given X at time $t=t_0$.

D. TLP

65. TLP is entitled “Tracking Loose-limbed People” and is a paper submitted for the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (“2004 CVPR”). In my opinion, TLP was widely disseminated and made publicly available by at least April 9, 2004.

66. TLP was presented at the 2004 CVPR for Poster Session 2. *See* EX1010 at 4. In fact, TLP was listed next to my own paper in the 2004 CVPR program. *Id.* at 4 (listing TLP next to “A Segmentation-Free Approach for Skeletonization of Gray-Scale Images via Anisotropic Vector Diffusion” by Z. Yu and C. Bajaj). When submitting a paper for the 2004 CVPR, paper submissions are done well in advance of the actual conference. For the 2004 CVPR, the deadline for full paper submission was November 19, 2003. EX1010 at 1. After submission,

the papers are widely disseminated and made publicly available to academics for peer review. After that process, authors are notified as to whether the papers have been accepted. For the 2004 CVPR, notification to the authors was March 2, 2004.

Id. For papers that are accepted, authors submit manuscripts for the conference.

The deadline for manuscript submission for the 2004 CVPR was April 9, 2004. *Id.*

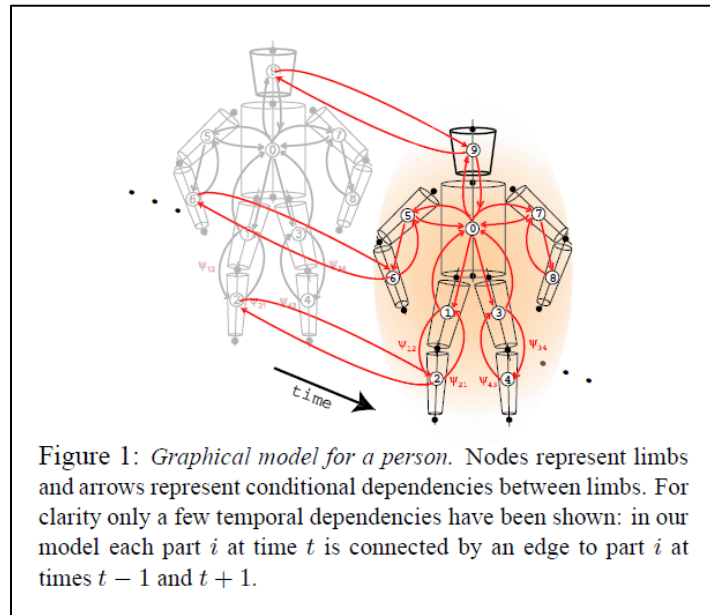
Since TLP was a paper presented at the 2004 CVPR, it is my opinion that it underwent the submission process outlined above, just as my own paper was.

Accordingly, it is my opinion that TLP was widely disseminated and made publicly available through the submission process, and no later than April 9, 2004.

67. In addition, I understand that TLP was disclosed during prosecution. EX1002 at 37-48. As such, TLP is cited as prior art on the face of the '980 Patent, and TLP is further cited in the provisional application that the '980 Patent claims priority to. EX1001 at 1 ("OTHER PUBLICATIONS ... Tracking loose-limbed people, CVPR, 2004"); EX1003 at 12 ("References ... Tracking Loose-limbed People, CVPR, 2004.")

68. TLP presents a probabilistic method for human detection and tracking in 3D. TLP uses a loose-limbed body model in which the limbs are connected via learned probabilistic constraints. Specifically, TLP describes a graphical model for detecting objects, such as a human, that comprise of nodes corresponding to the object's constituent parts. Edges in the graphical model correspond to position and

angle relationships between adjacent body parts in space and time, as illustrated in Figure 1:



EX1009 at 1.

69. Tracking is formulated as “one of inference in a graphical model and belief propagation is used to estimate the pose of the body at each time step.”

EX1009 at 1.

IX. DETAILED ANALYSIS

A. Ground I: Claims 1-16 Are Rendered Obvious Over Zhao

1. Claims 1/9

a) Element 1/9[preamble]

70. In my opinion, Zhao disclosed or at least rendered obvious this element. The preamble of Claims 1 and 9 are substantively similar, except Claim 1 recites “[a] computer implemented method for object detection” while Claim 9

recites “[a] computer readable medium embodying a computer program to perform method steps for object detection.” EX1001 at Claims 1, 9. In my opinion, both preambles were disclosed or at least rendered obvious in view of Zhao.

71. Zhao concerns object detection methods in the field of computer vision. EX1006 at 17 (“This dissertation presents an integrated human shape modeling, detection, and body part localization vision system.”); 97 (“This thesis has addressed several computer vision problems...”). As such, in my opinion a POSITA would have understood Zhao to describe a “computer implemented method” to perform object detection. Indeed, object detection is a field within image processing where computers are utilized to perform the object detection. Accordingly, methods for object detection in computer vision are naturally computer implemented methods.

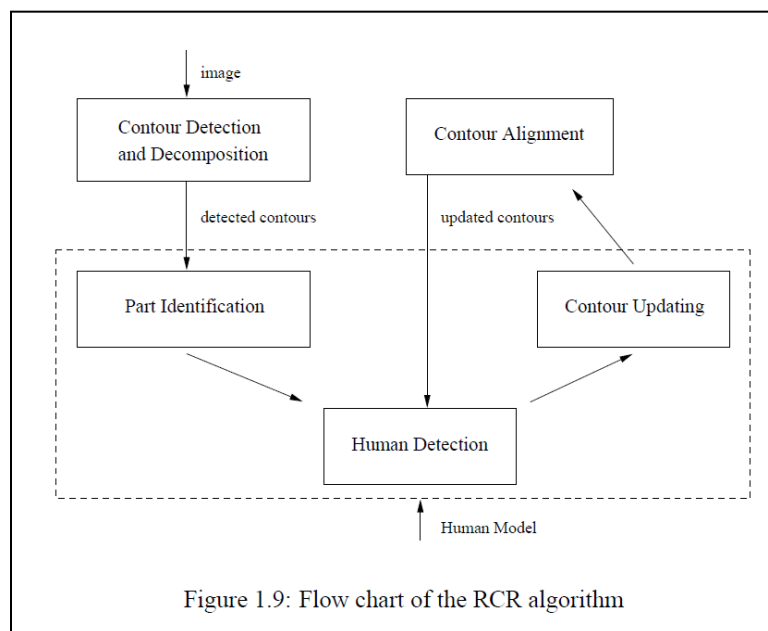
72. Further, as Zhao teaches a computer implemented method for object detection, in my opinion a POSITA would have found it obvious to have “a computer readable medium embodying a computer program” to perform the method for object detection. As a POSITA would know, computer implemented methods include the use of computer programs, which are instruction(s) that a computer executes. Computer programs are stored on “a computer readable medium” such as optical discs, cards, magnetic disks (i.e., hard drives), and other forms of computer storage. So, in my opinion, Zhao disclosed or at least rendered

obvious “a computer readable medium embodying a computer program” to perform the method for objection detection.

73. Accordingly, in my opinion, Zhao disclosed or at least rendered this element obvious.

b) Element 1/9[a]

74. In my opinion, Zhao disclosed or at least rendered this element obvious. As Zhao details for its object detection method, Zhao uses an object model in the form of a “human model” for a human to be detected. EX1006 at 17-18, 31 (“Human modeling is an essential part of model-based human detection.”). Zhao illustrates using this model in the flow chart of Figure 1.9:



EX1006 at 26.

75. In my opinion, the human object model described in Zhao—which includes spatial and temporal elements—disclosed or at least rendered obvious the claimed “spatio-temporal model.”

76. Regarding the “spatial” element, as an initial matter, I note that even the ’980 Patent does not contend that incorporating spatial elements into an object model is a point of novelty. Indeed, it was “common to represent objects as collections of features with distinctive appearance, spatial extent, and position.” EX1001 at 1:14-19. Further, in my opinion a POSITA would have found it obvious for a model for an object to be detected to include spatial elements. Indeed, as a POSITA would have known, *spatio*-temporal models were generally known in the prior art. As just one example, the paper by Sidenbladh (*see* EX1016) presents a Bayesian approach to tracking 3D articulated human figures using video image sequences. The human body is represented by spatial elements which are articulated cylinders viewed under perspective projection. A generative feature space video image model is defined in terms of the shape, appearance, and motion of the body, and a model of noise in the pixel intensities. This leads to a likelihood function that specifies the probability of observing components of objects in an image given the parameters of the human spatio-temporal model. A prior probability distribution over model parameters depends on the *temporal dynamics* of the body and the history of body shapes and motions. With this likelihood

function and temporal prior, the Sidenbladh paper estimates the posterior distribution over model parameters at each instant, given the observation history.

77. Further, specifically with respect to the Zhao reference, Zhao states that its “model is designed to encode the shapes of the body parts and the context information—the size and *spatial relationships* between body parts.” EX1006 at 69. Specifically, Zhao incorporates the spatial relationships between body parts as part of its model for a human to be detected through several model matrices: (1) an aspect ratio vector, (2) a length ratio matrix, (3) a relative position vector, and (4) an orientation or posture vector:

“Assuming that the human body model consists of m body parts that are parameterized with vectors $f_1, f_2, \dots, f_m$, where $f_i = (a_i, l_i, x_i, y_i, \theta_i)$. Then the human body model is parameterized with four model matrices: the aspect ratio vector $A = \{a_1, \dots, a_m\}$, the length ratio matrix $S = \{s_{ij}\}$, $i, j = 1, \dots, m$, where $s_{ij} = l_i/l_j$, the relative position vector $X = (x_1, y_1, \dots, x_m, y_m)$, and the orientation or posture vector $\Theta = \{\theta_1, \dots, \theta_m\}$.”

EX1006 at 43. Accordingly, it is my opinion that the “spatial” element of the claimed “spatio-temporal model” was disclosed or at least would have been obvious to a POSITA.

78. In my opinion, Zhao also disclosed or at least rendered obvious the claimed “spatio-*temporal* model” by incorporating temporal elements into its object model. According to Zhao, there are several prior art methods that include “motion-based” recognition procedures using “a sequence of images” for identification. EX1006 at 82; *see also* EX1003 at 5 (prior art methods for object

detection included “temporal propagation”). As such, a POSITA would have known to incorporate temporal elements into an object model for an object to be detected by incorporating information measured from multiple image frames, *i.e.*, information measured over time. *See, e.g.*, EX1011 at 1, 9 (“Generating Spatiotemporal Models from Examples” which details an approach for generating spatiotemporal models that “allow[] accurate temporal extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”).

79. Furthermore, probabilistic graphical model schemes have been effective for temporal tracking of image projections (appearances) in video, representing the presence of people wearing increasingly complex clothing in increasingly complex cluttered backgrounds. *See, e.g.*, EX1015, EX1028, EX1029, EX1041. Additionally, a paper titled “Attractive People” proposed a “loose-limbed” spatio-temporal body model in which the limbs are not rigidly connected but are rather “attracted” to each other (hence the title “Attractive People”). *See* EX1015. Instead of representing the body as a single 33-dimensional spatio-temporal graphical model, each limb is treated quasi-independently with soft constraints between the position and orientation of adjacent parts. The spatial constraints between body parts are defined as directed edges in the graph. Each edge has an associated conditional distribution that models the probabilistic

relationship between the parts. Each node in the graph also has a corresponding image likelihood function that models the probability of observing various image measurements conditioned on the position and orientation of the part. Person detection (or tracking) then exploits *belief propagation* (see EX1042) to estimate the belief distribution over the parameters which takes into account the constraints and the observations. This graphical inference problem can be carried out using any of the methods, Non-parametric Belief Propagation (NBP) (see EX1012) or the PAMPAS algorithm (see EX1019) that allows the parameters of the individual parts to be modeled using *continuous-valued* random variables rather than the discrete variables used in EX1035. A spatio-temporal graphical model is not new for finding or tracking articulated objects in video image frames and dates back at least to Fischler and Elschlager's pictorial structures. See EX1033. Variations on this type of model have been applied by Burl et al., Felzenszwalb and Huttenlocher, Coughlan and Ferreira, and Ramanan and Forsyth. See EX1034, 1035, EX1021, EX1036.

80. With respect to Zhao's disclosures, its "RCR algorithm is useful for vision tasks such as object tracking as well by incorporating motion information into the model." EX1006 at 99. Zhao extends the algorithm of sequential location of the projected video image appearances of articulated components of a deterministic object model to a TRS (Translation, Rotation, Scale) probabilistic

model and extends the method to update the shapes and sizes of the detected body parts of the humans, besides the positions of each of the body parts. EX1006 at 72-77 (citing EX1031). Zhao addresses issues of occlusion when detecting humans, for example, where overlapping body parts can create ambiguity and be confused with one another. To address this, “[m]otion information obtained from previous frames can be used to predict the orientations of the limbs and to solve the ambiguity.” EX1006 at 91. Figure 5.6 illustrates the application of this to a walking person:



EX1006 at 94. In the first half of the cycle no motion information is used “while in the second half of the cycle, the prediction from previous frames (constant angular

velocity is assumed) is used to get better results of body part localization.” EX1006 at 92.

81. In addition, Figure 5.5 illustrates the impact of incorporating temporal elements, such as motion information from previous frames, in Zhao’s system.

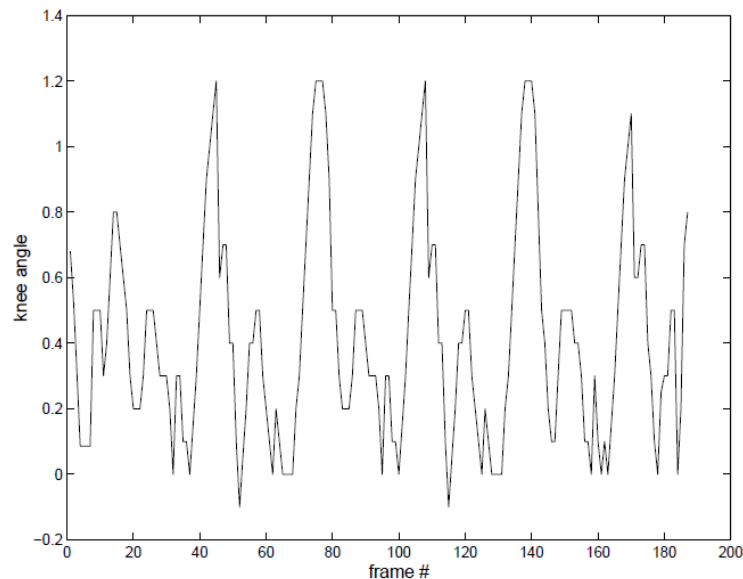


Figure 5.5: The change of angle at the knee with time

EX1006 at 93.

82. At the start, in frames 1-30, knee angle is not well observed because the left and right legs are sometimes confused with one another, but in frames 31-190 “an obvious pattern of walking cycle can be observed” because motion information is incorporated into the model (*i.e.*, frames 31-190), EX1006 at 92. (“[M]otion information is not used from frames 1 to 30, but is from frames 31 to 190.”). Accordingly, in my opinion, a POSITA would have found the claimed “spatio-temporal model for an object to be detected” disclosed or at least obvious

in view of Zhao because the Zhao model incorporates temporal elements for an object to be detected.

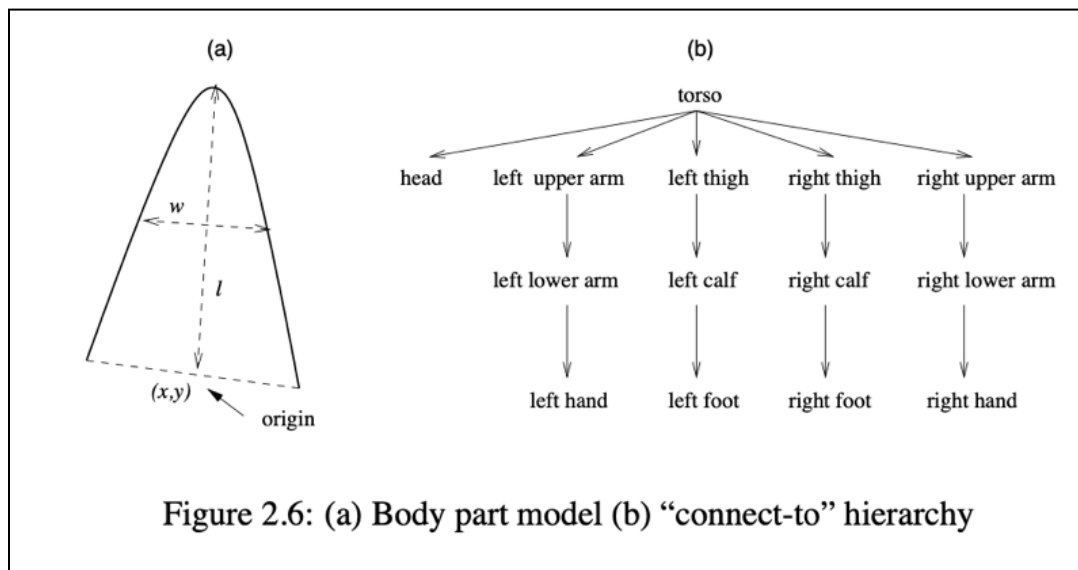
83. I understand that the Patent Owner may contend that the term “spatio-temporal model” should be construed as “a graphical model comprising nodes corresponding to each of the collection of components and to the object.” I offer no opinion on claim construction at this time, as Zhao disclosed or at least rendered obvious this element even under that purported construction. As a preliminary point, I believe a POSITA would have found the use of a graphical model in computer vision and object detection obvious, as graphical models were known in the prior art long before the ’980 Patent. *See, e.g.*, EX1006, 52 (“Pope and Lowe [prior art] modeled an object with a graph whose nodes represent the feature values and whose edges represent the spatial arrangement (symmetric, parallel) of the features.”). This is recognized by the ’980 Patent itself, which does not purport to invent graphical models. EX1003 at 5 (“We should note that the use of graphical models and belief propagation is not novel for object detection.”). In fact, such models were widespread in the prior art before the ’980 patent. Furthermore, probabilistic schemes using spatio-temporal graphical models have been effective for temporal tracking of image projections in video, representing the belief of the presence of people wearing increasingly complex clothing in increasingly complex cluttered backgrounds. *See* EX1015, EX1028, EX1029, EX1041. Additionally, the

paper titled “Attractive People” uses a “loose-limbed” spatio-temporal body model in which the limbs are not rigidly connected but are rather “attracted” to each other (hence the title “Attractive People”). *See* EX1015. Instead of representing the body as a single 33-dimensional spatio-temporal model, each limb is treated quasi-independently with soft constraints between the position and orientation of adjacent parts. The spatial constraints between body parts are defined as directed edges in the graph. Each edge has an associated conditional distribution that models the probabilistic relationship between the parts. Each node in the graph also has a corresponding image likelihood function that models the probability of observing various image measurements conditioned on the position and orientation of the part. Person detection (or tracking) then exploits *belief propagation* (*see* EX1042) to estimate the belief distribution over the parameters which takes into account the constraints and the observations. This graphical inference problem can be carried out using any of the methods, Non-parametric Belief Propagation (NBP) (*see* EX1012) or the PAMPAS algorithm (*see* EX1019) that allows the parameters of the individual parts to be modeled using *continuous-valued* random variables rather than the discrete variables used in EX1035. A spatio-temporal graphical model is not new for finding or tracking articulated objects in video image frames and dates back at least to Fischler and Elschlager’s pictorial structures. *See* EX1033. Variations on this type of model have been applied by Burl et. al.,

Felzenszwalb and Huttenlocher, Coughlan and Ferreira, and Ramanan and Forsyth.

See EX1034, EX1035, EX1021, EX1036.

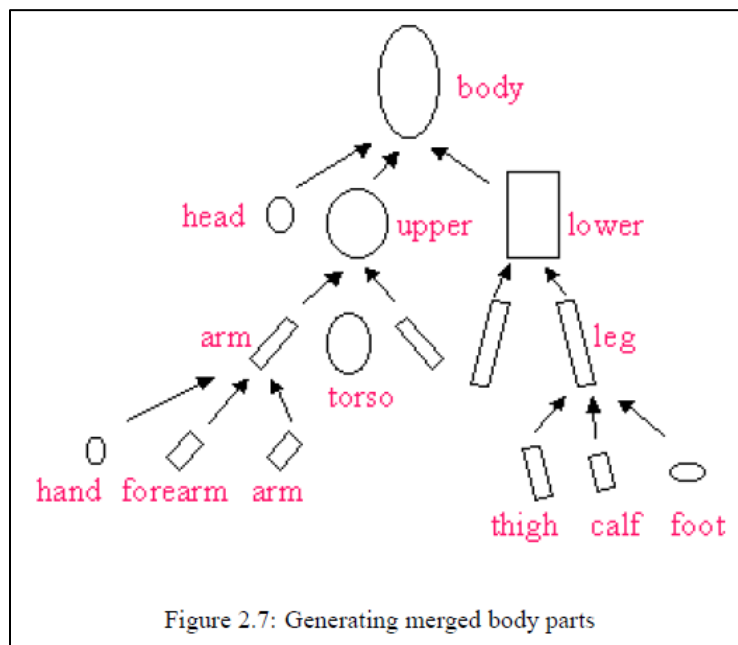
84. Further, in my opinion a POSITA would have found this element disclosed or at least obvious, as the Zhao object model is a graphical model with nodes corresponding to an object and a collection of its components. Zhao explains, “[e]ach human body model consists of six main body parts (the head, the torso, two arms and two legs) and twelve subparts of the four limbs. The body parts are constrained to connect to each other at the joints.” EX1006 at 42. The model is defined as a graphical model that’s comprised of “ribbons” for these body parts, as shown in Figure 2.6(a):



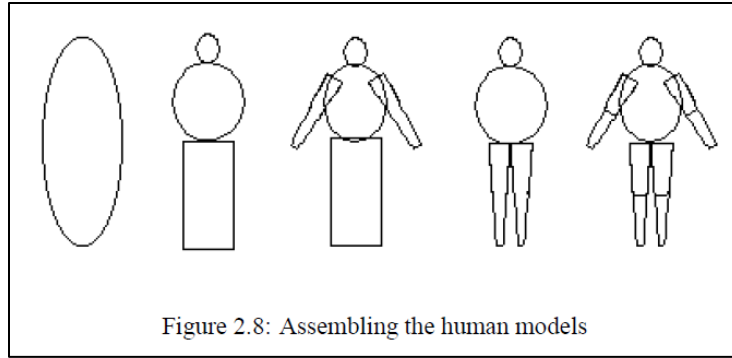
EX1006 at 42 (“The body parts are modeled with ribbons as shown in Fig. 2.6(a)”) (citations omitted). As illustrated above, the ribbons are graphical representations of

body parts and are “nodes” corresponding to each of the collection of components and to the object.

85. Further, in my opinion the Zhao object model incorporates a “coarse-to-fine modeling of human shapes” that merges body parts. These “merged parts are also modeled with ribbons.” EX1006 at 46. Figure 2.7 of Zhao illustrates this procedure:



EX1006 at 45. As such, Zhao’s teachings can handle the challenging problem of “shape variation from clothing” that “may drastically change the appearance of a person.” EX1006 at 45-46. Accordingly, Zhao’s object model, which is a “coarse-to-fine” model, includes “levels between the whole body [*i.e.*, the object] and the major body parts to include the merged body parts [*i.e.*, the collection of components],” as illustrated in Figure 2.8 of Zhao:



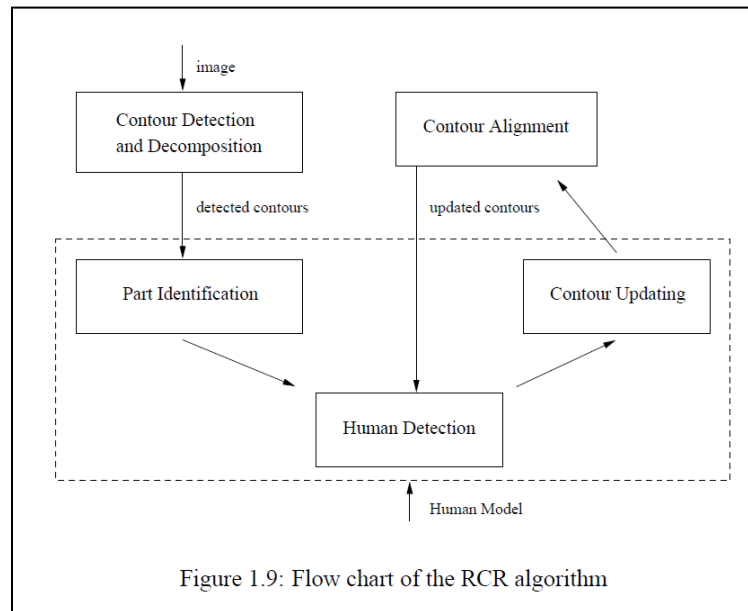
EX1006 at 46, 48 (“all parts are organized in a hierarchy to facilitate coarse-to-fine shape decomposition and classification”).

86. Accordingly, in my opinion, a POSITA would have found the purported spatio-temporal model disclosed or at least obvious in view of Zhao, as the “ribbons” serve as “nodes” that correspond to each of the collection of components (*i.e.*, ribbons corresponding to component body parts) and the object (*i.e.*, ribbons corresponding to merged parts).

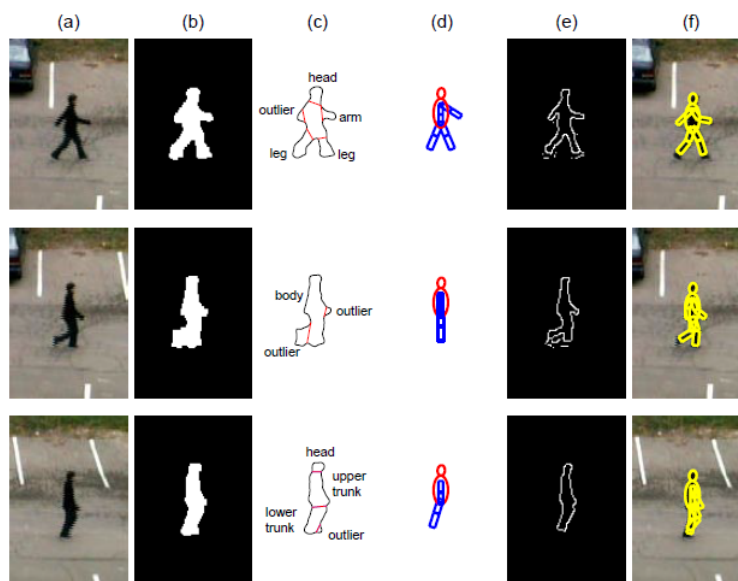
c) Element 1/9[b]

87. In my opinion, Zhao disclosed or at least rendered obvious this this limitation. Zhao describes a “human shape modeling, detection, and body part localization vision system,” which is used on image data including in videos. EX1006 at 17, 41 (“the goal of this thesis is to detect people in an image”), 62 (depicting people and body parts detected in a “raw image.”). Zhao applies its system to a “pedestrian detection system” that “has been tested on the videos of urban areas obtained from a stereo system mounted on the top of a minivan.” EX1006 at 84-85. A POSITA would have understood a video to comprise a

plurality of images, as a video is by definition simply an inter-related sequence of images. Figure 1.9 illustrates providing video image data (*i.e.*, a plurality of images) that includes the object:



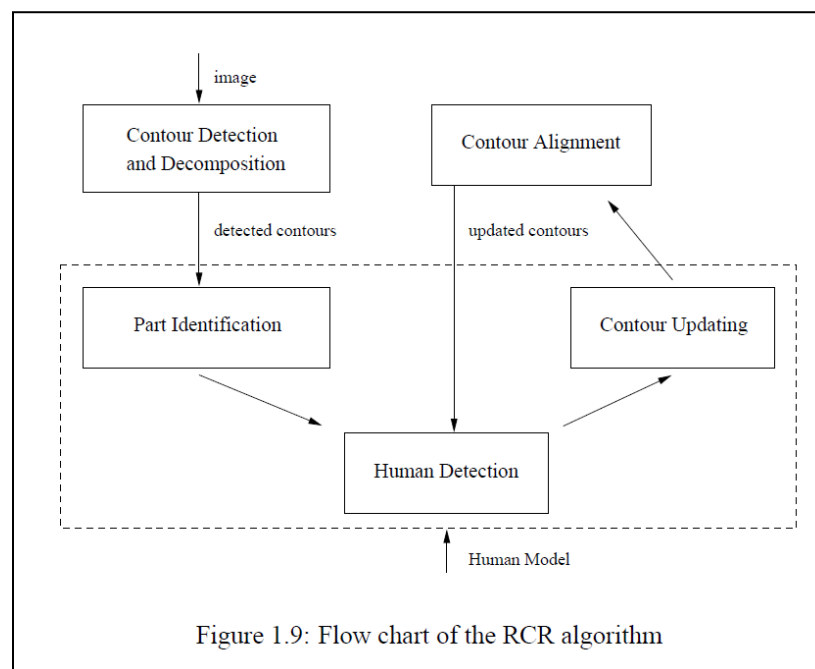
EX1006 at 26. Also, Figure 5.4 illustrates the application of Zhao on a a series of image frames (in other words, a video) with a person walking:



Id at 92. Accordingly, in my opinion Zhao disclosed or at least rendered this limitation obvious.

d) Element 1/9[c]

88. In my opinion, Zhao disclosed or at least rendered this limitation obvious. To measure the object in each image, Zhao begins with “contour detection and decomposition” as illustrated in the flow chart of Figure 1.9:



EX1006 at 26. Contour detection extracts the silhouettes of objects and is “commonly used to overcome variable texture” and “many contour extraction methods have been proposed” in the prior art. EX1006 at 20. After extraction, contours or silhouettes are decomposed into component parts, as illustrated in Figure 2.4:

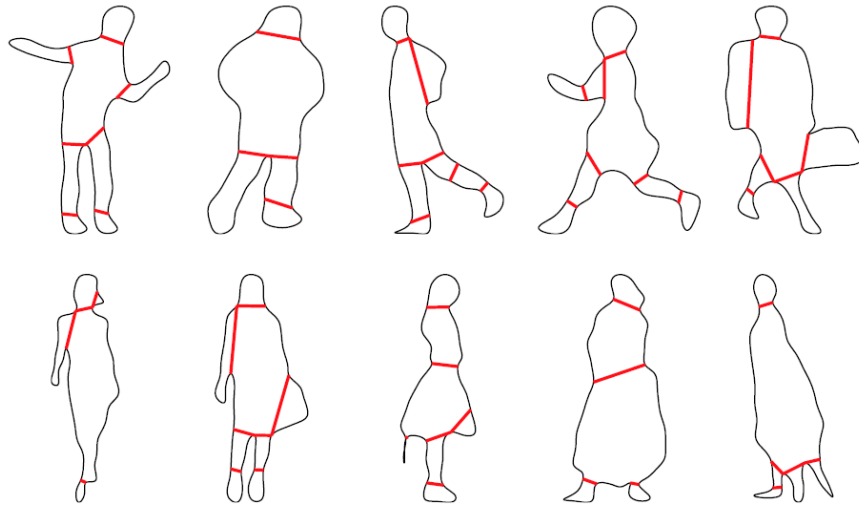


Figure 2.4: Example results of natural shape decomposition

EX1006 at 40. Zhao then performs the steps of part identification and human detection. Accordingly, Zhao measures the object as “a collection of components.”

89. Furthermore, Zhao uses a “coarse-to-fine search procedure to identify and locate the body parts” (EX1006 at 57) with components having an associated state vector defining the component’s position and scale within an image. *See* EX1006 at 43 (“In summary, a body part is parameterized with a vector (a, l, x, y, θ) , where $a = w / l$ is the aspect ratio that captures the general shape of a ribbon, (x, y) are the coordinates of the origin in the coordinate frame of the parent part, and θ is the intersection angle between the major axes of this part and its parent part.”). This matches the process described by the ’980 Patent specification, where nodes representing an object’s components “have an associated state vector $X^T=(x,y,s)$

defining the component's real-valued position and scale within an image.” EX1001 at 3:50-52.

90. Accordingly, in my opinion a POSITA would have found the element disclosed or at least rendered obvious.

e) Element 1/9[d]

91. In my opinion, Zhao disclosed this limitation or at least rendered it obvious. Zhou explains, “[a] TRS-invariant probabilistic model is designed to encode the shapes of the body parts and the context information — the size and spatial relationships between body parts. A Bayesian framework is developed to perform human detection and part identification under partial occlusion. A contour updating procedure is introduced to integrate the human model and the identified body parts to predict the shapes and locations of the parts missed by the contour detector; the refined contours are used to re-evaluate the Bayesian similarity measure and determine if the detected contour is a person or not. Therefore, contour extraction, body part localization, and human detection are improved by combining the context constraints from the identified body parts and the human model.” EX1006 at 69.

92. After measuring an object as a collection of components, Zhao determines the probability that the object is in each image. To do so, Zhao uses a “Bayesian Similarity Measure,” which is described as a “part-based similarity

measure [that] evaluates the resemblance between a contour and a model based on the best match between their body parts.” EX1006 at 54.

93. Zhao starts with “[a] simple decision rule to perform human detection” where the probability the “contour C corresponds to a person” in each image is calculated:

$$P(\text{person}|C) > \text{threshold}$$

EX1006 at 64. By applying Bayes’ theorem and conditioning the hypothesis for body part identification, determining the probability that a contour corresponds to a person (i.e., determining a probability that the object is in each image) is re-formulated as:

$$P(C|H^*, \text{person}) P(\text{person}|H^*)$$

where H^* is known as “the maximum *a posteriori* (MAP) hypothesis ... selected from the hypothesis space”, is the optimal or best hypothesis of the optimization

$$H^* = \arg \max_H P(C|H, \text{person}) P(\text{person}|H).$$

such that a given person is present in the image.

EX1006 at 54.

94. The *goodness function* that rates the hypothesis is defined as follows:

$G(H) \triangleq P(C|H, \text{person}) P(\text{person}|H)$. Here, “ $P(C|H, \text{person})$ ” evaluates the degree of resemblance between the matched pairs, while $P(\text{person}|H)$ is proportional to the number of identified body parts — the more body parts being identified, the more

likely the extracted contour is a person. Thus, the MAP hypothesis H^* selected using [the above optimization,] maximizes the resemblance between the matched pairs and the number of identified body parts.” EX 1006 at 54-55. Zhao explains “[m]ost previous methods [EX1006: References 78, 79, 86] detect body parts sequentially without backtracking. The disadvantage is that if one of the body parts is located incorrectly, the locations of the rest of the body parts would be wrong.” EX1006 at 57. In contrast, Zhao selects the optimal hypothesis through dynamic model assembling “every extracted ribbon is considered as a candidate for every body part. The exhaustive search for the best hypothesis H^* is to check every possible hypothesis and find the best one based on the goodness function $G(H)$.” EX1006 at 57.

95. As formulated by Zhao, $P(C|H, person)$ represents “the degree of resemblance between the matched pairs.” EX1006 at 55. “The likelihood $P(C|H, person)$ derived above constrains the best match between the set of decomposed ribbons and the model body parts to be of similar relative sizes and positions as well as similar aspect ratios.” EX1006 at 56. $P(person|H)$ is “proportional to the number of identified body parts—the more body parts being identified, the more likely the extracted contour is a person.” EX1006 at 55. “The conditional probability $P(person|H)$ is estimated based on the number and types of the identified body parts. The more body parts being identified, the more likely the

extracted contour is a person. And the presence of a fine level body part indicates a higher likelihood of the presence of a person than that of a coarse level body part does.” EX1006 at 56. “[T]he following formula is designed to estimate $P(\text{person}|H)$ ”:

$$P(\text{person}|H) = \sum d_i w_i \quad (3.7)$$

EX1006 at 56.

96. Additionally, I will note that probabilistic frameworks for determining a probability that an object and which of its components are present in an image were widespread and obvious to a POSITA. In computer vision applications such as video human tracking, graphical models employ approximate inference algorithms such as Bayesian optimization (BO) and loopy belief propagation (BP) used in Alspach et. al. and Isard (*see* EX1020 and EX1019) to infer the position and scale of the object and components. The BO and BP approximate inference methods produce good results for a wide range of articulate models as also shown in papers by Coughlan et. al. and De la Torre. *See* EX1021 and EX1022. For temporal inference problems defined on Markov chains, and temporal graphical models, a variety of spatio-temporal object feature density representations have been explored over a moving window of video image frames. These probabilistic methods include Gaussian sum filters (*see* EX1020) that use local linearizations to deterministically propagate Gaussian mixture models in space and time through the

video image sequence frames. In contrast, particle filters (*see* EX1024) use Monte Carlo methods to stochastically update a set of weighted point samples of feature density representations of probabilistic graphical models within the temporal window of video image sequences.

97. Accordingly, in my opinion, a POSITA would have recognized that Zhao renders obvious the element “determining a probability that the object is in each image.”

f) Element 1/9[e]

98. In my opinion, Zhao disclosed this limitation or at least render it obvious. In addition to determining the probability that the object is in each image, Zhao compares the probability to a threshold to perform object detection. Zhao formulates the “human detection problem” as “given a detected contour C , determine whether this contour represents the silhouette of a person or not.” EX1006 at 64. Zhao begins with a simple decision rule that compares the probability of an extracted contour matching a person with a threshold: “the contour C corresponds to a person if

$$P(\text{person}|C) > \text{threshold}.$$

Otherwise, C is not a person.” EX1006 at 64. By applying Bayes’ formula and conditioning the hypothesis for body part identification, Zhao reformulates the decision rule as: “the contour C corresponds to a person if

$$P(C|H^*, person)P(person|H^*) > threshold$$

EX1006 at 65.

99. Notably, employing a threshold in a probability measure was known to a POSITA, as a “threshold” is a basic, fundamental part of providing a Boolean result (such as “detected” vs. “not detected”) that would be used in a probabilistic measure to perform any type of detection, and there are many examples of using thresholds in probabilistic frameworks. *See, e.g.,* EX1038 at 15 (Alspach) (“experimental evidence shows that specific choices of number of terms in the *a priori* densities and specific thresholds for the two approximations lead to a bound on the number of terms required to describe the *a posteriori* densities for any stage”); EX1041 at 3 (Sidenbladh, Black) (“...the likelihood of observing the filter responses for an image is proportional to the ratio between the likelihood that the foreground image pixels are explained by the foreground object and the likelihood that they are explained by some general background” . . . “This ratio is highest when the foreground (person) model projects to an image region that is unlikely to have been generated by some general scene but is well explained by the statistics of people.”); EX 1017, EX1018; EX1040 at (Viola and Jones) (explicitly disclosing that the probabilistic thresholds are experimentally set using techniques such as ADABOOST: “The key insight is that smaller, and therefore more efficient, boosted classifiers can be constructed which reject many of the negative sub-

windows while detecting almost all positive instances (i.e. the threshold of a boosted classifier can be adjusted so that the false negative rate is close to zero). ... Stages in the cascade are constructed by training classifiers using AdaBoost and then adjusting the threshold to minimize false negatives.”).

100. Accordingly, in my opinion, a POSITA would have understood that Zhao renders this limitation obvious and discloses “detecting the object in any image upon comparing the probabilities for each image to a threshold for detecting the object.”

2. Claims 2/10

101. In my opinion, Zhao disclosed or at least rendered obvious Claims 2/10. Claims 2 and 10 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or at least rendered Claims 1/9 obvious. My analysis of Claims 1/9 is incorporated fully herein by reference.

102. As described above for limitation 1/9(a), Zhao discloses a “spatio-temporal model” for an object to be detected. In my opinion, Zhao further teaches the use of detectors for each of the collection of components. For example, Zhao explains that its algorithm matches decomposed body parts against a “set of model body parts” to detect body parts (*i.e.*, the collection of components). EX1006 at 54. “ $P(C|H, person)$ evaluates the degree of resemblance between the matched pairs, while $P(person|H)$ is proportional to the number of identified body parts—the

more body parts being identified, the more likely the extracted contour is a person.” EX1006 at 55. Accordingly, a POSITA would have understood that Zhao disclosed or at least rendered obvious a spatio-temporal model for an object to be detected that is comprised of “detectors” where each of the collection of components (*i.e.*, body parts) is matched with a “set of model body parts.”

103. As such, in my opinion, Claims 2/10 was disclosed or at least rendered obvious.

3. Claims 3/11

104. In my opinion, Zhao disclosed or at least rendered obvious Claims. Claims 3 and 11 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or at least rendered obvious Claims 1/9. My analysis of Claims 1/9 is incorporated fully herein by reference.

105. Further, the additional requirement of Claim 3/11—“wherein the spatio-temporal model is a graphical model comprising nodes corresponding to each of the collection of components and to the object”—matches the Patent Owner’s purported construction of the term “spatio-temporal model.” Accordingly, Claim 3/11 is disclosed or at least rendered obvious for the same reasons discussed for limitation 1/9(a) and said purported construction. As described above, I believe a POSITA would have found the purported spatio-temporal model disclosed or at least obvious in view of Zhao, as it describes a graphical model where “ribbons”

serve as “nodes” that correspond to each of the collection of components (*i.e.*, the component body parts) and the object (*i.e.*, the merged parts forming a person). My analysis of of limitation 1/9(a) is incorporated fully herein by reference.

4. Claims 4/12

106. In my opinion, Zhao rendered Claims 4/12 obvious. Claims 4 and 12 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or rendered obvious Claims 1/9. My analysis of Claims 1/9 is incorporated fully herein by reference.

107. As described above for Claims 1/9, Zhao determines the probability that an object is in each image based on identified body parts. However, Zhao states that in detecting objects, overlapping body parts can create ambiguity and be confused with one another. So, “[m]otion information obtained from previous frames can be used to predict the orientations of the limbs and to solve the ambiguity.” EX1006 at 91. Accordingly, in my opinion, a POSITA would have understood Zhao to teach a method of detecting an object in a current image according to measurements of the object as a collection of components from other image frames, including prior and later image frames.

108. For example, with Zhao, the first step is to integrate the *identified* body parts and the human model to refine the probability estimates of the locations of the identified body parts. The second step is to predict the outlines of the

unidentified body parts from the human model and the identified body parts. This is done by a MAP (maximum *a posteriori*) probability likelihood estimation of a body part's parameter vector corresponding to the detection outlines in the image. The location of the main body part is estimated from the locations of the other main body parts in the image frames and the motion information obtained from the images. If more than two main body parts have been MAP identified, then the TRS (translation, rotation, scaling) transformation that transforms the coordinates from the spatio-temporal human model coordinate system to the image coordinate system is also calculated as part of the coarse-to-fine backtracking search procedure. Zhao explicitly discloses the best MAP probability estimates that an object is in each image by its coarse-to-fine, backtracking search procedure is based on the shapes and locations of the identified, and the predicted body parts through use of the spatio-temporal human model, to enable measurements of the human object as a collection of components determined from a prior image and a later image relative to the current image.

109. Notably, analyzing data in image frames from video by using measurements from adjacent frames, including previous and later frames, was widespread and commonly understood in image processing. *See* EX1011 at 1, 9 (“Generating Spatiotemporal Models from Examples” which details an approach for generating spatio-temporal models that “allow[] accurate temporal

extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”); *see also* EX1037 (Andrew Blake & Michael Isard, “Active Contours”, Springer-Verlag (1998)) (describing a framework and analysis for Bayesian inference of visual shape and motion from video); EX1024 at 339-357, Chapter 16 (Andrew Blake, Michael Isard & John MacCormick, “*Statistical Models of Visual Shape and Motion*,” in A. Doucet, N. de Freitas, & N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York (2001)). Accordingly, while Zhao explicitly describes this principle by detecting an object according to measurements of the object as a collection of components determined from a “prior image,” a POSITA would have further understood that doing so with a “later image” to be obvious.

110. As such, in my opinion, Claims 4/12 would have been obvious over Zhao.

5. Claims 5/13

111. In my opinion, Zhao disclosed or at last rendered obvious Claims 5/13. Claims 5 and 13 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or at least rendered obvious Claims 1/9. My analysis of Claims 1/9 is incorporated fully herein by reference.

112. As described above for limitation 1/9(a), and in Claim 4/12, Zhao describes a spatio-temporal model for an object to be detected and tracked. In detecting objects such as humans, problems with occlusion arise such that overlapping body parts can create ambiguity and be confused with one another. So, Zhao explains that “[m]otion information obtained from previous frames can be used to predict the orientations of the limbs and to solve the ambiguity.” EX1006 at 91. In view of Zhao’s teachings to look to measurements from “previous frames” a POSITA would have understood it to be obvious to define a “temporal window” to do so. Indeed, a POSITA would have known that a widespread and common technique in image processing was to define a temporal window of image frames to analyze data. Using measurements in that temporal window for analyzing the current image was obvious and understood. *See* EX1011 at 1, 9 (“Generating Spatiotemporal Models from Examples” which details an approach for generating spatiotemporal models that “allow[] accurate temporal extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”); EX1037 (describing a framework and analysis for Bayesian inference of visual shape and motion from video); EX1024, Chapter 16.

113. Further, in view of Zhao, a POSITA would have understood that measurements of the object’s components within a temporal window (*i.e.*, such as

a temporal window of the “previous frames” described in Zhao) are passed to components detected in the current image. Figure 5.5 illustrates the effect of passing measurements of components detected in the temporal window:

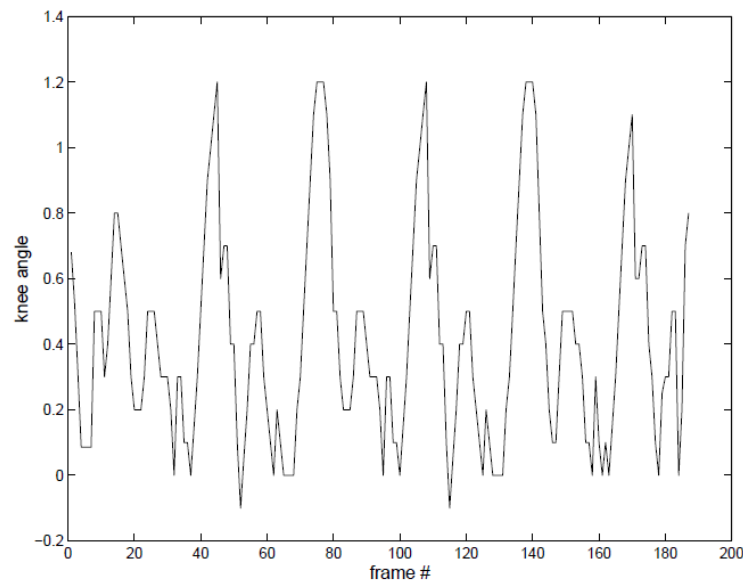
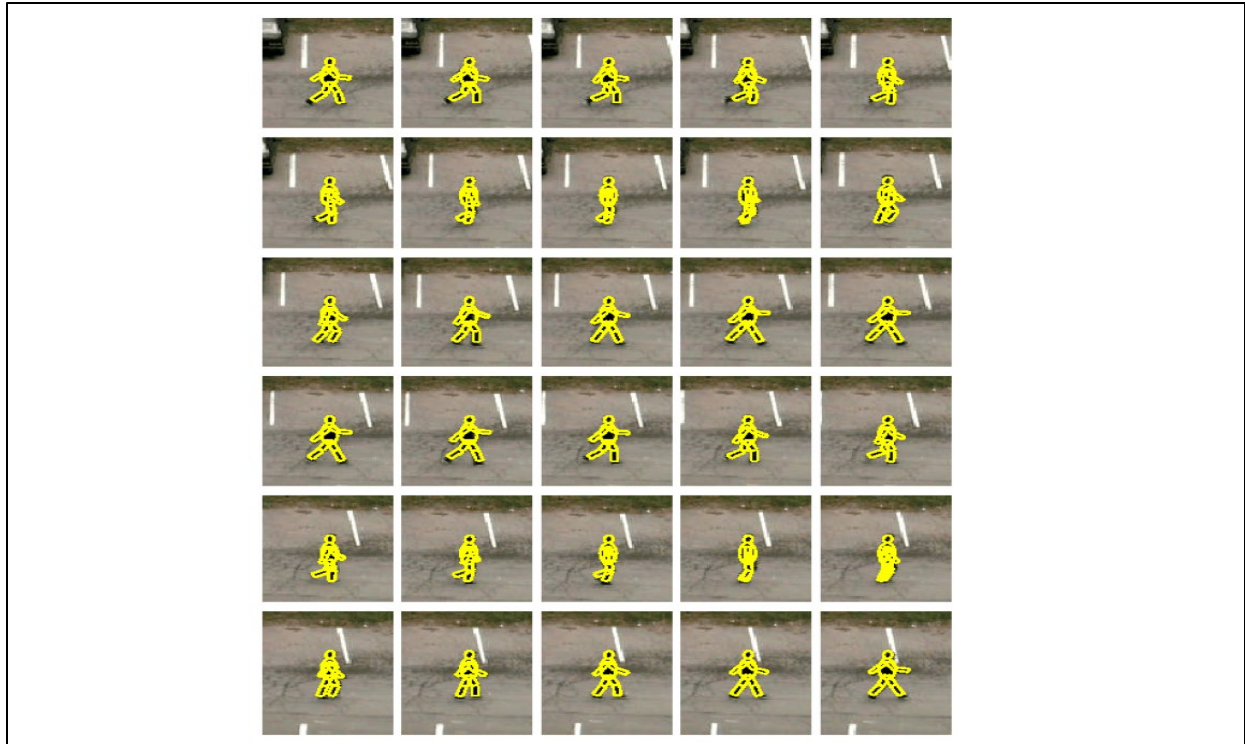


Figure 5.5: The change of angle at the knee with time

EX1006 at 93.

114. At the start in frames 1-30, knee angle is not well observed because the left and right legs are sometimes confused, but in the later frames 31-190 “an obvious pattern of walking cycle can be observed” after incorporating motion information into the model. EX1006 at 91-92 (“[M]otion information is not used from frames 1 to 30, but is from frames 31 to 190.”). As another example, Figure 5.6 illustrates the effect of passing measurements of components detected in a temporal window:



EX1006 at 94. No motion information is used in the first half of the cycle “while in the second half of the cycle, the prediction from previous frames (constant angular velocity is assumed) is used to get better results of body part localization.” EX1006 at 92. As such, measurements of the body parts from the temporal window (*i.e.*, the “prediction from previous frames”) are passed to to the components detected in the current frame.

115. Accordingly, in my opinion, Claims 5/13 were disclosed or at least rendered obvious by Zhao.

6. Claims 6/14

116. In my opinion, Zhao disclosed or at least rendered obvious Claims 6/14. Claims 6 and 14 depend on Claims 1 and 9, respectively. As described above,

Zhao disclosed or at least rendered Claims 1/9 obvious. My analysis of Claims 1/9 is incorporated fully herein by reference.

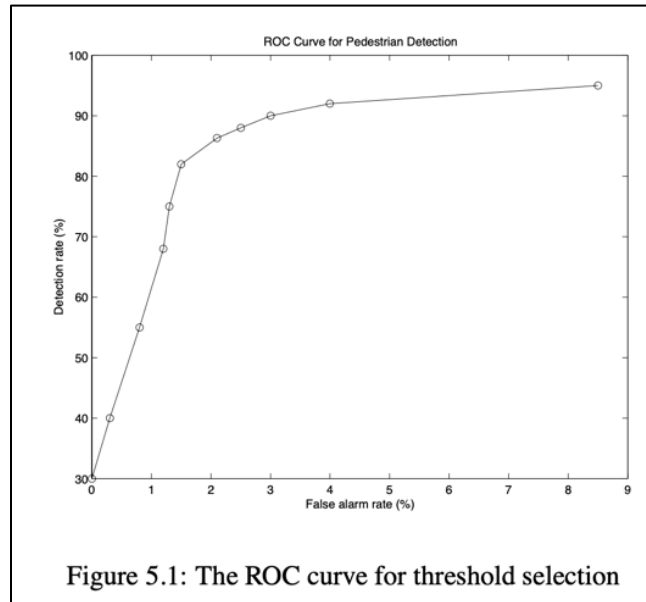
117. As described above for element 1/9(d), Zhao describes determining the probability that an object is in each image. Zhao further states, “[t]he above probability distributions provide metrics to evaluate the *shape, size relationship, and configuration* similarities between the detected contour and the human body model.” EX1006 at 45. For example, the probability distribution “parameters (means and covariances) are estimated from the measurements provided by Tilley[116].” *Id.* (citing EX1039, Alvin R. Tilley, “The Measure of Man and Woman: Human Factors in Design,” H.D. Associates, NY (1993)). In my opinion, A POSITA would have understood Zhao to evaluate the shape, size relationship, and configuration, which is a determination of a probability for a position and a size of the object in each image as required by claims 6/14.

7. Claims 7/15

118. In my opinion, Zhao disclosed or at least rendered obvious Claims 7/15. Claims 7 and 15 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or at least rendered obvious Claims 1/9. My analysis of Claims 1/9 is incorporated fully herein by reference.

119. As described above for limitation 1/9(e), Zhao uses probabilistic framework for object detection that includes the use of a threshold. Additionally,

the threshold is selected empirically based on an “ROC curve” and the rate of false detections are “generated by testing the RCR algorithm on 20 sequences (100 frames/sequences) of urban street scenes,” as illustrated in Figure 5.1:



EX1006, 84-85. A POSITA would have understood that the threshold is determined “empirically” because the threshold is based off the rate of false detections, which was measured by applying and testing the system. Similar examples of empirical threshold adjustments are explained in EX1040 (Paul Viola & Michael Jones “Rapid Object Detection using a Boosted Cascade of Simple Features”, Conference on Computer Vision and Pattern Recognition) (2001). Notably, a POSITA would have understood that “empirical” results are determined through experimentation or observation, instead of only theoretical or arbitrary measures. Accordingly, in my opinion, a POSITA would have understood that Zhao describes that the threshold may be determined empirically.

8. Claims 8/16

120. In my opinion, Zhao rendered Claims 8/16 obvious. Claims 8 and 16 depend on Claims 1 and 9, respectively. As described above, Zhao disclosed or at least rendered Claims 1/9 obvious. My analysis of Claims 1/9 is incorporated fully herein by reference.

121. As described for elements 1/9(d)-(e), Zhao presents a probabilistic framework for a spatio-temporal model. Accordingly, the joint probability used in Zhao, and well known to a POSITA, is easily representable in the form of the joint probability distribution formulated by Claims 8/16. In Claims 8/16, the equation represents a typical factorized form of the joint probability distribution $P(\dots)$ for a graphical model G of an object O with state X_i^O at time frame i of the video and $\{k = 1..N\}$ components state X_i^{C-k} . This form of the joint probability would have been well known to a POSITA as a pairwise Markov field representation of a two-level spatio-temporal graphical model, the first level representing the object state at each time frame, and the second level for individual components states, and captured by their pairwise probabilistic relationships as defined by edges in the graphical model. See EX1012 at 1 (§ 2.1).

122. Accordingly, in my opinion, Zhao rendered Claims 8/16 obvious.

B. Ground II: Claims 1-16 Are Obvious Over Steffens In View Of Zhao

1. Motivation To Combine

123. In my opinion, a POSITA would have been motivated to combine, and would have had a reasonable expectation of success, in combining Steffens and Zhao. As an initial matter, Steffens and Zhao both concern computer vision and teach methods for object detection in video images. EX1007, 1:14-17; EX1006, 17 (“This dissertation presents an integrated human shape modeling, detection, and body part localization vision system.”); 97 (“This thesis has addressed several computer vision problems...”). In my opinion, a POSITA would have naturally looked to and combined their teachings, which are complementary, because Steffens and Zhao each similarly concern object detection and recognition in video images based on multi-component object models, which are expressed as spatio-temporal graphical models. Given the substantial similarities between the methods and algorithms in Steffens and Zhao, a POSITA would have had a reasonable expectation of success in combining their teachings.

124. Notably, Zhao explains that its method for object detecting as applied to humans is broadly applicable and “will also shed light on general objection detection and shape analysis.” EX1006 at 19. As such, while Steffens focuses its object detection method on facial recognition, a POSITA would have been motivated to look to combine the teachings of Zhao and Steffens to complete the

full picture of detecting a person and their facial expressions for a more fulsome computer vision of the person. A POSITA would have had a reasonable expectation of success in doing so since the teachings of Zhao and Steffens were interchangeably applicable, regardless of whether a face or a human body (or other object) was the subject of the detection.

125. As part of its method for object detection, Steffens describes “identifying the object based on a similarity” between an image and a corresponding model. EX1007 at Claim 13, 7:63-67 (“The search for optimum similarity may be performed by positioning the model graph in the upper left hand corner of the image, extracting the jets at the nodes, and determining the similarity between the image graph and the model graph.”). While Steffens says that “confidence values” . . . “indicate the reliability of face detection and which are generated during bunch graph matching,” it does not provide all details on how an object is identified based on a similarity measure. EX1007 at 5:12-13. But as a POSITA would have known and appreciated, similarity measures may include probabilistic frameworks for identifying objects in video images and recognizing their motions. Zhao expressly describes a “similarity measure” in the form of a “Bayesian Similarity Measure,” which “combines the local and global relationship constraints into a single equation to evaluate the degree of resemblance.” EX1006 at 25. Like Steffens, Zhao describes a method where an object is identified based

on a similarity between an image and a corresponding model (*i.e.*, “the resemblance between a contour [of an object in an image] and a model”). EX1006 at 54. Not only would a POSITA have been motivated to combine Steffens and Zhao, a POSITA would have had a reasonable expectation of success applying Zhao’s probabilistic Bayesian Similarity Measure to Steffens, since Zhao is elaborating on and improving a similarity measure, which Steffens already incorporates into its method for object detection. Indeed, A POSITA would have been further motivated to combine Steffens with Zhao because “the Bayesian similarity measure enables efficient shape matching and comparison robust to articulation, partial occlusion, and segmentation errors through coarse-to-fine human model assembling.” EX1006 at 98.

126. Accordingly, in my opinion, a POSITA would have been motivated to combine the teachings of Steffens and Zhao and would have a reasonable expectation of success in doing so.

2. Claims 1/9

a) Element 1/9[preamble]

127. In my opinion, this element was disclosed or at least would have been obvious over Steffens in view of Zhao. As I describe above for Ground I, this

element was disclosed or at least would have been rendered obvious in view of Zhao. My previous analysis is incorporated by reference.

128. In my opinion, this element was also disclosed or at least rendered obvious in further view of Steffens, as its “present invention relates to vision-based object detection and tracking, and more particularly, to systems for detecting objects in video images, such as human faces, and tracking and identifying objects in real time.” EX1007 at 1:14-17. Steffens is describing a system and method in computer vision for detecting objects in video images. As such, POSITA would have understood Steffens to describe a computer implemented method for object detection. As a computer implemented method, POSITA would have understood this to further include “a computer readable medium embodying a computer program” to perform that method.

b) Element 1/9[a]

129. In my opinion, this element was disclosed or at least would have been rendered obvious over Steffens in view of Zhao. As I describe above for Ground I, this element was disclosed or at least would have been rendered obvious in view of Zhao. My previous analysis is incorporated by reference.

130. In my opinion, this element also was disclosed or at least would have been rendered obvious in further view of Steffens. Steffens provides a model for an object to be detected, which is a “spatio-temporal model.”

131. Regarding the “spatio” element, for its model for an object to be detected, Steffens says that “the nodes of the labeled graph refer to points on the object and are labeled by jets,” and “[e]dges of the graph are labeled with distance vectors between the nodes.” EX1007 at 8:58-60. Further, “[n]odes and edges define the graph topology. Graphs with equal geometry may be compared.” EX1007 at 60-62. Accordingly, a POSITA would have understood that “spatial” elements are used for an object to be measured. For example, by incorporating the “distance vectors between the nodes,” the model is incorporating spatial information for an object to be detected. As such, Steffens disclosed the “spatio” aspect of the spatio-temporal model or at least rendered it obvious.

132. For the “temporal” element, Steffens explains that when “the object is in a sequence of images . . . the step of detecting an object further includes tracking the object between image frames based on a trajectory associated with the object. Also, the step of locating the nodes includes tracking the nodes between image frames and reinitializing a tracked node if the node's position deviates beyond a predetermined position constraint between image frames.” EX1007 at 2:6-13. As part of its process for recognizing objects, Steffens claims “tracking the object between image frames based on a trajectory associated with the object,” “tracking the nodes between image frames,” and “reinitializing a tracked node if the node’s position deviates beyond a predetermined position constraint between image

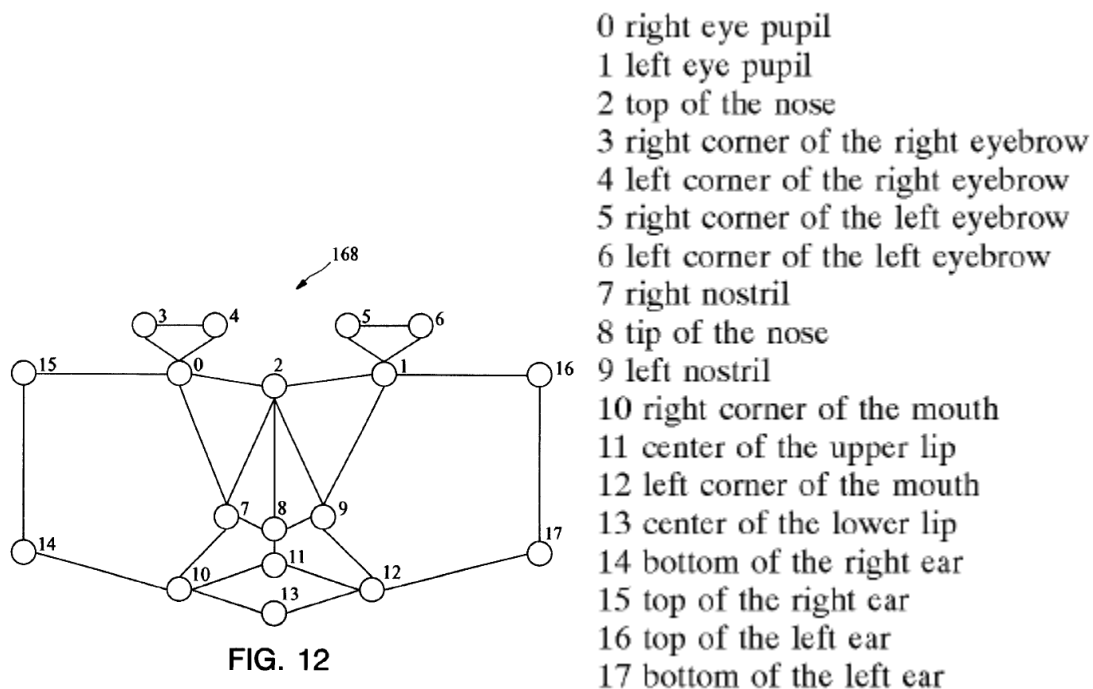
frames.” *See, e.g.*, EX1007 at Claims 14, 16, and 17. Accordingly, in my opinion, a POSITA would have understood that Steffens describes a model for an object to be detected that includes using “temporal” information, such as the location of an object and its nodes between image frames. As such, in my opinion, Steffens disclosed or at least rendered obvious the “temporal” element of the spatio-temporal model.

133. Further, to the extent the term “spatio-temporal model” is construed as “a graphical model comprising nodes corresponding to each of the collection of components and to the object,” I believe Steffens still disclosed or at least rendered this element obvious. As a preliminary matter, using graphical models for object detection was known well before the ’980 Patent, as a POSITA would have known and appreciated. *See supra*, Ground I, limitation 1(a). In any event, Steffens renders the spatio-temporal model obvious, as it also describes a graphical model with nodes corresponding to an object and a collection of its components.

134. According to Steffens, it uses an object model to perform “bunch graph matching” and conduct similarity comparisons to identify an object. Like an image graph, a bunch graph is a data structure that “is used to describe the aspects of an object.” EX1007 at 8:56-57, 9:22-25. The graph includes “nodes” that correspond to points on an object that are labeled with “jets.” For a bunch graph “instead of attaching only a single jet to each node, a whole bunch of jets 172 (a

bunch jet) are attached to each node.” EX1007 at 9:22-25. Jets are derived from facial images and to form a bunch graph, “a collection of facial images (the bunch graph gallery) is marked with node locations at defined positions of the head,” which are called “landmarks.” EX1007 at 9:25-29. Accordingly, Steffens describes providing a “graphical model comprising nodes corresponding to each of the collection of components and to the object.”

135. A model graph that is designed for finding a human face is illustrated in Figure 12 with nodes corresponding to a collection of components:



EX1007 at 13 (Fig. 12).

136. To improve computational efficiency, Steffens teaches a “coarse-to-fine” approach using fewer nodes. EX1007 at 15:32-41. After coarse landmark finding has been achieved, higher precision localization may be performed on

higher resolution images for precise finding of a particular facial feature. An example of the coarse-to-fine procedure is illustrated by Figure 14:

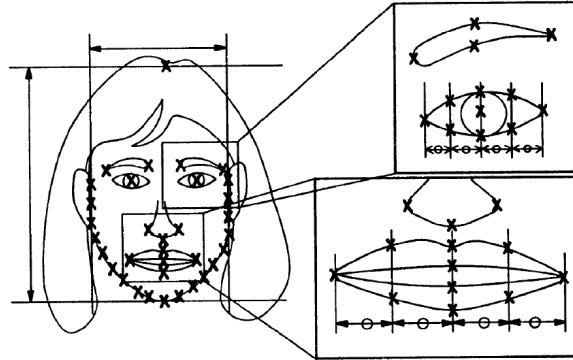


FIG. 14

EX1007 at 15 (Fig. 14).

137. As illustrated above, Steffens' "coarse" approach uses fewer nodes that correspond to the object and its components, where for example, one node is used to represent an eye. Steffens' "fine" approach uses more nodes, including for example several nodes outlining the eye and a node at the center of the pupil. Accordingly, in my opinion, providing a spatio-temporal model comprised of "nodes" corresponding to each of the collection of components and to the object (*i.e.*, a coarse-to-fine approach) was disclosed or at least would have been obvious to a POSITA.

c) Element 1/9[b]

138. In my opinion, this element was disclosed or at least would have been obvious over Steffens in view of Zhao. As I describe above for Ground I, this

element was disclosed or at least would have been obvious in view of Zhao. My previous analysis is incorporated by reference.

139. In my opinion, this element was further disclosed or at least rendered obvious over Steffens. Steffens states, “[t]he present invention relates to vision-based object detection and tracking, and more particularly, to systems for detecting objects in *video images*, such as human faces...” EX1007 at 1:14-16. Accordingly, “[t]he object recognition process 10 operates on digitized video image data provided by an image processing system 12. The image data includes an image of an object class, such as a human face. The image data may be a single video image frame or a series of sequential monocular or stereo image frames.” EX1007 at 3:48-53.

140. Accordingly, in my opinion, a POSITA would have understood that Steffens teaches providing a video comprising a plurality of images including the object, as its object detection method “operates on digitized video image data provided by an image processing system” where “[t]he image data includes an image of an object class, such as a human face.” *Id.* Accordingly, Steffens disclosed or at least rendered this limitation obvious.

d) Element 1/9[c]

141. In my opinion, this element was disclosed or at least would have been obvious over Steffens in view of Zhao. As I describe above for Ground I, this

element was disclosed or at least would have been rendered obvious in view of Zhao. My previous analysis is incorporated by reference.

142. In my opinion, this element was further disclosed or at least rendered obvious over Steffens. As I explained for element 1/9(a), Steffens uses a graph matching technique for measuring an object. In doing so, “[t]he nodes 164 of the labeled graph refer to points on the object and are labeled by jets 160. Edges 166 of the graph are labeled with distance vectors between the nodes. Nodes and edges define the graph topology” EX1007 at 8:58-61. Steffens’ disclosures match the process described in the ’980 Patent specification, where nodes representing an object’s components “have an associated state vector $X^T=(x,y,s)$ defining the component’s real-valued position and scale within an image.” EX1001 at 3:50-52. Like the ’980 Patent specification, which describes nodes with an associated state vector to measure an object as a collection of components, Steffens uses nodes with an associated state vector (*e.g.*, distance vectors between the nodes) to measure the object as a collection of components. *Id.*

143. Accordingly, in my opinion, a POSITA would have found this element disclosed or at least rendered obvious.

e) Element 1/9[d]

144. In my opinion, this limitation was disclosed or at least would have been rendered obvious in view of Steffens in combination with Zhao. Steffens

describes its method for object detection as using a “similarity” measure. Although not explicitly formulated as calculating a “probability” that an object is an image, a POSITA would have found doing so to be obvious. Indeed, probabilistic frameworks for object detection were widely known in the prior art and frequently used as part of tracking objects in image video sequences. A typical approach was a *Bayesian* approach. The human body is represented by articulated cylinders and viewed under perspective projection. A spatio-temporal model is defined in terms of the shape, appearance, and motion of the body, and a model of noise in the pixel intensities. This leads to a likelihood function that specifies the probability of observing an image given the model parameters. A prior probability distribution over model parameters depends on the temporal dynamics of the body and the graphical model of linked body component shapes and motions. With this likelihood function and temporal prior, one formulates the posterior distribution over model parameters for the joint probability that the object (level 1) and each of its constituent component parts (level 2) is visible in the image and tracked across image frames. *See also, supra*, Ground I, element 1/9(d).

145. Accordingly, in my opinion, a POSITA would have been motivated to combine the teachings of Steffens with a probabilistic framework, such as Zhao. Steffens and Zhao both use a similarity measure for the identification of objects in video data. As explained in Ground I, for Zhao, this includes the use of a

“Bayesian Similarity Measure” that determines the probability that an object is in each image of a video. Accordingly, in my opinion, for the reasons described above for Zhao, “determining a probability that the object is in each image” was disclosed or at least would have been rendered obvious to a POSITA.

f) Element 1/9[e]

146. In my opinion, this element was disclosed or at least would have been obvious in view of Steffens and in combination with Zhao. As explained above for element 1/9(d), it would have been obvious to a POSITA to use a probabilistic framework, which was widely known in the prior art. Further, as part of a probabilistic framework, it would have been obvious to a POSITA to detect the image by comparing the probabilities for each image to a threshold for detecting the object.

147. Notably, employing a threshold in a probability measure was known to a POSITA, as a “threshold” is a basic, fundamental part of providing a Boolean result (such as “detected” vs. “not detected”) that would be used in a probabilistic measure to perform any type of detection, and there are many examples of using thresholds in probabilistic frameworks. *See, e.g.,* EX1038 at 15 (Alspach) (“experimental evidence shows that specific choices of number of terms in the a priori densities and specific thresholds for the two approximations lead to a bound on the number of terms required to describe the a posteriori densities for any

stage”); EX1041 at 3 (Sidenbladh, Black) (“...the likelihood of observing the filter responses for an image is proportional to the ratio between the likelihood that the foreground image pixels are explained by the foreground object and the likelihood that they are explained by some general background” . . . “ This ratio is highest when the foreground (person) model projects to an image region that is unlikely to have been generated by some general scene but is well explained by the statistics of people.”); EX1017; EX1018; EX1040 at 4 (Viola and Jones) (explicitly disclosing that the probabilistic thresholds are experimentally set using techniques such as ADABOOST: “The key insight is that smaller, and therefore more efficient, boosted classifiers can be constructed which reject many of the negative sub-windows while detecting almost all positive instances (i.e. the threshold of a boosted classifier can be adjusted so that the false negative rate is close to zero). . . . Stages in the cascade are constructed by training classifiers using AdaBoost and then adjusting the threshold to minimize false negatives.”).

148. Moreover, as explained above, in my opinion, a POSITA would have been motivated to look for a probabilistic framework, such as Zhao, to combine with the teachings of Steffens. Both Steffens and Zhao use a similarity measure for the identification of objects in video data. As explained in Ground I, for Zhao, this includes the use of a “Bayesian Similarity Measure” that determines the probability that an object is in each image of a video. This further includes comparing the

probabilities for each image to a threshold for detecting an object. *See* Ground I limitation 1/9(e).

149. Accordingly, in my opinion, for the reasons described above for Zhao, this element was disclosed or at least would have been obvious to a POSITA.

3. Claims 2/10

150. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 2/10. Claims 2 and 10 depend on Claims 1 and 9, respectively. As described above, Steffens in combination with Zhao disclosed or at least renders Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 2/10. *See, supra*, Ground I, Claims 2/10. The analysis for Ground I is incorporated herein by reference.

151. Further, as described above, Steffens discloses a “spatio-temporal model” for an object to be detected. Steffens further teaches the use of detectors for each of the collection of components. For example, Steffens explains that its process uses a graph matching technique where the graphs include “nodes” that refer to points on the object, which are labeled with “jets.” EX1007 at 8:58-59. Objects are “identified based on a similarity between wavelet jets associated with an object image in a gallery of object images” and “wavelet jets at the nodes on the transformed image.” Notably, the model described in Steffens for an object to be

detected includes both spatial and temporal elements. For example, Steffens explains that a model graph for identifying an object (*i.e.*, a human face) contain nodes at various positions and “[e]dges of the graph are labeled with distance vectors between the nodes.” *Id.* at 8:59–61. Steffens further explains that detecting the object in video includes “tracking the object between image frames based on a trajectory associated with the object” and locating the nodes that correspond to an object’s features “includes tracking the nodes between image frames.” *Id.* at 2:8–12, and moreover that “the system of their invention is particularly advantageous for recognizing a person over a wide variety of pose angles” *Id.* at 3:44-46. “To compute the similarity between two graphs, the sum is taken over similarities of corresponding jets between the graphs.” EX1007 at 8:65-67. Accordingly, a POSITA would have found a spatio-temporal model for an object to be detected to be comprised of “detectors” to be disclosed or at least rendered obvious, where each of the collection of components (*i.e.*, parts of the object) is matched with the “jets.”

152. As such, in my opinion, Claims 2/10 was disclosed or at least would have been rendered obvious to a POSITA.

4. Claims 3/11

153. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 3/11, which depend on Claims 1/9, respectively. As

described above, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 3/11. *See, supra*, Ground I, Claims 3/11. The analysis for Ground I is incorporated herein by reference.

154. Further, the additional requirement of Claims 3/11 “wherein the spatio-temporal model is a graphical model comprising nodes corresponding to each of the collection of components and to the object” matches the Patent Owner’s purported construction of the term “spatio-temporal model.” Accordingly, Claims 3/11 were disclosed or at least rendered obvious for the same reasons discussed for limitation 1/9(a) and that purported construction. Indeed, Steffens describes a graphical model (i.e., a bunch graph) comprised of nodes (i.e., the nodes with jets) corresponding to each of the collection of components and to the object (*i.e.*, a coarse-to-fine approach where the nodes correspond to more or fewer object features for computational efficiency). *See supra*, limitation 1/9(a).

5. Claims 4/12

155. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 4/12, which depend on Claims 1/9, respectively. As described above, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by

reference. Further, as explained above for Ground I, Zhao rendered obvious Claims 4/12. *See, supra*, Ground I, Claims 4/12. The analysis for Ground I is incorporated herein by reference.

156. As described above in limitation 1(d), I believe it was disclosed or at it would have been obvious in view of Steffens in combination with Zhao to determine the probability that an object is in each image. Further, it was disclosed or at least it would have been obvious to a POSITA to detect an object in an image “according to measurements of the object as a collection of components determined from a prior image and a later image relative to the current image.” EX1001, Claim 4. As Steffens explains, when “the object is in a sequence of images ... the step of detecting an object further includes *tracking the object between image frames* based on a trajectory associated with the object. Also, the step of locating the nodes includes *tracking the nodes between image frames* and reinitializing a tracked node if the node's position deviates beyond a predetermined position constraint *between image frames*.” EX1007 at 2:6-13. Accordingly, Steffens claims as part of its process for recognizing objects, “tracking the object between image frames based on a trajectory associated with the object,” “tracking the nodes between image frames,” and “reinitializing a tracked node if the node’s position deviates beyond a predetermined position constraint between image frames.” *See, e.g.*, EX1007 at Claims 14, 16, and 17.

157. By tracking the object and/or its associated nodes “between image frames” as part of its process for recognizing objects, a POSITA would have found it disclosed or at least obvious to detect the object according to measurements of the object (which is measured as a collection of components, as discussed for limitation 1/9(a)) from a “prior image” and a “later image” relative to the current image. Indeed, a POSITA would have found it obvious that tracking an object and/or its associated nodes “between frames” would include doing so according to a “prior image” and a “later image” relative to the current image.

158. Notably, analyzing data in image frames from video by using measurements from adjacent frames, including previous and later frames, was widespread and commonly understood in image processing. *See* EX1011 at 1, 9 (“Generating Spatiotemporal Models from Examples” which details an approach for generating spatio-temporal models that “allow[] accurate temporal extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”); *see also* EX1037, (describing a framework and analysis for Bayesian inference of visual shape and motion from video); EX1024 at Chapter 16.

159. Accordingly, in my opinion, Claims 4/12 was disclosed or at least would have been obvious in view of Steffens in combination with Zhao.

6. Claims 5/13

160. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 5/13, which depend on Claims 1/9, respectively. As described above, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 5/13. *See, supra*, Ground I, Claims 5/13. The analysis for Ground I is incorporated herein by reference.

161. As described above for limitation 1/9(a), Steffens describes a spatio-temporal model for an object to be detected. When “the object is in a sequence of images ... the step of detecting an object further includes *tracking the object between image frames* based on a trajectory associated with the object. Also, the step of locating the nodes includes *tracking the nodes between image frames* and reinitializing a tracked node if the node's position deviates beyond a predetermined position constraint *between image frames*.” EX1007 at 2:6-13. Accordingly, as part of its process for recognizing objects, Steffens claims “tracking the object between image frames based on a trajectory associated with the object,” “tracking the nodes between image frames,” and “reinitializing a tracked node if the node’s position deviates beyond a predetermined position constraint between image frames.” *See, e.g.*, EX1007 at Claims 14, 16, and 17. A POSITA would have

understood to use a “temporal window” defining a plurality of images in which measurements of components detected are passed to components detected in the current image, as a POSITA would need to define a window of “image frames” for tracking the object and/or its associated nodes “between image frames.” Steffens explains that a model graph for identifying an object (*i.e.*, a human face) contain nodes at various positions and “[e]dges of the graph are labeled with distance vectors between the nodes.” *Id.* at 8:59–61. Steffens further explains that detecting the object in video includes “tracking the object between image frames based on a trajectory associated with the object” and locating the nodes that correspond to an object’s features “includes tracking the nodes between image frames.” *Id.* at 2:8-12, and moreover that “the system of their invention is particularly advantageous for recognizing a person over a wide variety of pose angles” *Id.* at 3:44-46. For example, Steffens describes a temporal window in an embodiment that includes two consecutive image frames, where “[t]he similarity (or confidence) value S may be calculated to indicate how well the two image regions in the two image frames correspond to each other...” EX1007 at 13:22-25.

162. Notably, a POSITA would have known that a widespread and common technique in image processing was to define a temporal window of image frames to analyze data. Using measurements in that temporal window for analyzing the current image was obvious and understood. *See* EX1011 at 1, 9

(“Generating Spatiotemporal Models from Examples” which details an approach for generating spatiotemporal models that “allow[] accurate temporal extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”); EX1037 (describing a framework and analysis for Bayesian inference of visual shape and motion from video); EX1038 at Chap. 16.

163. As such, in my opinion, Claims 5/13 was disclosed or at least would have been obvious in view of Steffens in combination with Zhao.

7. Claims 6/14

164. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 6/14, which depend on Claims 1/9, respectively. As described above, Steffens in combination with Zhao disclosed or at least rendered Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference.

165. As described above in limitation 1/9(d), Steffens in combination with Zhao describes determining the probability that an object is in each image. Further, Steffens states that its matching technique, includes for example, “provid[ing] more accurate information about the position and size of the face.” EX1007 at 7:18-20. Zhao also states, “[t]he above probability distributions provide metrics to evaluate the *shape, size relationship, and configuration* similarities between the detected contour and the human body model.” EX1006 at 45. By providing

information on position and size (*e.g.*, of the face as in Steffens) or the shape, size, relationship and configuration (*e.g.*, of the detected body as in Zhao), a POSITA would have found it obvious to determine the probability for a “position and size of the object” as claimed in the ’980 Patent. For example, and as indicated earlier, the probability distribution “parameters (means and covariances) of the position and size of the object can be estimated from the measurements provided by Tilley[116].” *Id.* (citing EX1039). Accordingly, Claims 6/14 was disclosed or at least would have been rendered obvious.

8. Claims 7/15

166. In my opinion, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 7/15, which depend on Claims 1/9, respectively. As described above, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference.

167. As described above for limitation 1/9(e), Zhao uses a threshold as part of its probabilistic framework for detecting an object. Zhao additionally explains that this threshold can be selected empirically based on an “ROC curve” and the rate of false detections “generated by testing the RCR algorithm on 20 sequences (100 frames/sequences) of urban street scenes,” as explained above for Ground I. *See, supra*, Ground I, Claims 7/15. In my opinion, Claims 7/15 was disclosed or at

least would have been obvious in view of Steffens in combination with Zhao for the same reasons.

9. Claims 8/16

168. In my opinion, Steffens in combination with Zhao rendered obvious Claims 8/16, which depend on Claims 1/9, respectively. As described above, Steffens in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference.

169. As described for Claims 1/9 limitations (d-e), using a probabilistic framework for a spatio-temporal model was disclosed or at least would have been obvious. Further, as described above in Ground I, the joint probability used in Zhao, and well known to a POSITA is easily representable in the form of the joint probability distribution formulated by Claims 8/16. *See supra*, Ground I, Claims 8/16. In my opinion, Claims 8/16 would have been obvious in view of Steffens in combination with Zhao for the same reasons.

C. Ground III: Claims 1-16 Are Obvious Over Ozer In View Of Zhao

1. Motivation To Combine

170. In my opinion, a POSITA would have been motivated to combine, and would have had a reasonable expectation of success, in combining Ozer and Zhao. As an initial matter, Ozer and Zhao both concern computer vision and teach methods for object detection in video images. EX1008 at Abstract (“the

methodology is able to identify object parts and to decide on the presence of the object of interest (human, bag, dog, etc.) in video sequences”); EX1006 at 17 (“This dissertation presents an integrated human shape modeling, detection, and body part localization vision system.”), 97 (“This thesis has addressed several computer vision problems...”). Because Ozer and Zhao each similarly concern spatio-temporal object models and measuring the object as a collection of image components, a POSITA would have naturally looked to and combined their complementary teachings, which a POSITA would have had a reasonable expectation of success. Indeed, both Ozer and Zhao provide spatio-temporal object models and measure the object as a collection of image components to detect objects such as humans and classifying them over time.

171. I understand that during prosecution, the examiner found that there was no motivation or suggestion to combine Ozer with the Cham prior art reference “because the object detection algorithms in [the] two references are vastly different.” EX1002 at 163. While Ozer described a graph matching algorithm that used a spatio-temporal model, Cham described a “spatial registration framework” and a kinematic model, not a spatio-temporal model that incorporated temporal elements into the object detection algorithm. *Id.* (citing Cham).

172. But unlike the Cham reference considered during prosecution, in my opinion, Ozer and Zhao both similarly describe methods and algorithms for object detection, so a POSITA would have been motivated to combine them. Zhao and Ozer each compare extracted portions of an image with a set of stored models, such as graphical models. EX1008 at 7:8-31. Both references also describe the use of a spatio-temporal model for an object to be detected and are not limited to the “kinematic model” described in Cham.

173. Further, when Ozer describes, for example, “matching the relational graphs of objects with [a] reference model” (EX1008 at 4:24-25), Ozer explains that the graph matching algorithm looks for “meaningful correspondence” between the object parts and the reference model. *Id.* at 7:8-31. As a POSITA would have known, finding “meaningful correspondence” through graph matching may include using probabilistic frameworks to perform similarity analyses. In computer vision applications such as video human tracking, probabilistic methods based on graphical models employ approximate inference algorithms such as Bayesian optimization (BO) and loopy belief propagation (BP) used in Alspach et. al. and Isard (*see* EX1020 and EX1019) to infer the position and scale of the object and components. The BO and BP approximate inference methods produce good results for a wide range of articulate models as also shown in papers by Coughlan et. al. and De la Torre. *See* EX1021 and 1022. For temporal inference problems defined

on Markov chains, and temporal graphical models, a variety of spatio-temporal object feature density representations have been explored. These include Gaussian sum filters (*see* EX1020) that use local linearizations to deterministically propagate Gaussian mixture models in space and time. In contrast, particle filters (*see* EX1024) use Monte Carlo methods to stochastically update a set of weighted point samples of feature density representations of probabilistic graphical models. One such framework is expressly described in Zhao as a “Bayesian Similarity Measure,” which “combines the local shape and global relationship constraints into a single equation to evaluate the degree of resemblance.” EX1006 at 25. Like Ozer’s search for “meaningful correspondence,” Zhao describes a method where an object is identified based on a similarity between an image and a corresponding model (*i.e.*, “the resemblance between a contour [of an object in an image] and a model”). EX1006 at 54. As such, a POSITA would have been motivated to combine Ozer and Zhao and had a reasonable expectation of success applying Zhao’s probabilistic Bayesian Similarity Measure to Ozer, since Zhao is elaborating on and improving a process for finding meaningful correspondence (or similarity), which Ozer already incorporates into its method for object detection. A POSITA would have been further motivated to combine Ozer with Zhao’s Bayesian similarity measure, as “the Bayesian similarity measure enables efficient shape matching and comparison robust to articulation, partial occlusion, and

segmentation errors through coarse-to-fine human model assembling.” EX1006 at 98.

174. Accordingly, in my opinion, a POSITA would have been motivated to combine the teachings of Ozer and Zhao.

2. Claims 1/9

a) Element 1/9[preamble]

175. In my opinion, this element was disclosed or at least rendered obvious over Ozer in view of Zhao. As an initial matter, this element was disclosed or at least would have been obvious in view of the teachings of Zhao, as described above for Ground I. The analysis for this element under Ground I is incorporated herein by reference.

176. Further, Ozer describes a method for object detection in video sequences. EX1008 at Abstract. As a system and method relating to object detection for video images, a POSITA would have understood Ozer to describe a computer implemented method for object detection. A POSITA would have understood this to further include “a computer readable medium embodying a computer program” to perform that method.

b) Element 1/9[a]

177. In my opinion, this element was disclosed or at least would have been rendered obvious over Ozer in view of Zhao. As an initial matter, this element was disclosed or at least would have been obvious in view of the teachings of Zhao, as

described above for Ground I. The analysis for this element under Ground I is incorporated herein by reference.

178. Further, in my opinion, this element was disclosed or at least would have been obvious in view of the teachings of Ozer. Ozer describes providing a “spatio-temporal model” for an object to be detected. Indeed, Ozer describes a process for “graph matching” where “we compare the object model with a set of stored models. Each extracted region modeled with ellipses corresponds to a node in the graphical representation of the object of interest.” EX1008 at 7:8-11. Further, Ozer teaches “Classifying Over Multiple Frames” where “[t]he movements of the object parts are described as a spatio-temporal sequence of feature vectors that consist of the direction of the object part movement. The system checks direction of the movements of the object parts for a number of frames and calculates the probabilities of the activities with the known activities by using Hidden Markov Models and chooses the pattern with the highest probability as the recognized activity in these frames (408).” EX1008 at 7:33-45, 4:45-48 (describing the use of “Hidden Markov Models, which classifies activity patterns of the body parts in space-time and determines the movements of the object parts. It also uses a distance classifier to detect specific gestures and activities.”). Accordingly, in my opinion, a POSITA would have found the claimed “spatio-temporal model” disclosed or at least obvious in view of Ozer.

179. Further, to the extent the term “spatio-temporal model” is construed as “a graphical model comprising nodes corresponding to each of the collection of components and to the object,” in my opinion, Ozer disclosed or at least rendered this element obvious. As part of Ozer’s method for object detection, each extracted part in an image “corresponds to a node in the graphical representation of the object of interest.” EX1008 at 7:9-11. The “[o]utput of the graph-matching algorithm is the classified object parts.” *Id.* at 7:33-35. “For example, if the object of interest is a human body, the system will first detect different object parts, e.g., hands, head, arms, legs, torso and compare these part attributes with the human model attributes via graph matching.” *Id.* at 4:38-41. Further, “geometrical characteristics of the object regions is used to group object parts” and “[t]he object is decomposed for its presentation as a combination of component shapes,” such that “[t]he result will be unaffected by a partial occlusion of the object.” *Id.* at 4:1-6. Accordingly, in my opinion, a POSITA would have understood Ozer to describe providing a graphical model comprising nodes [i.e., the graphical representation of the object of interest] corresponding to each of the collection of components [i.e., the classified object parts] and to the object [i.e., the grouped object parts, forming the object].

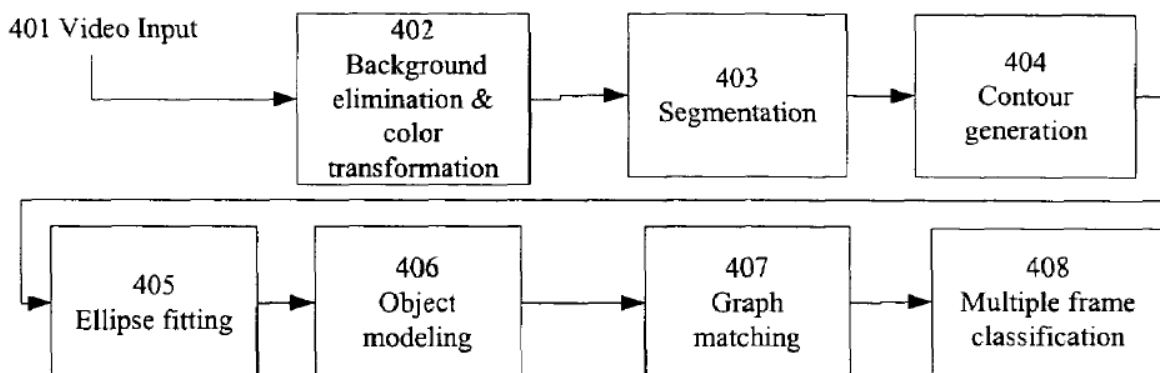
180. As such, in my opinion, this limitation was disclosed or at least would have been obvious.

c) Element 1/9[b]

181. In my opinion, this element was disclosed or at least would have been rendered obvious over Ozer in view of Zhao. As an initial matter, this element was disclosed or at least would have been obvious in view of the teachings of Zhao, as described above for Ground I. The analysis for this element under Ground I is incorporated herein by reference.

182. In my opinion, Ozer further disclosed this limitation or at least rendered it obvious. Ozer states that its “methodology is able to identify object parts and to decide on the presence of the object of interest (human, bag, dog, etc.) in video sequences.” EX1008 at Abstract. Accordingly, Ozer’s system provides “401 Video Input” to perform object detection, as illustrated in Figure 4:

Figure 4



EX1008, 6 (Fig. 4).

183. Moreover, as Ozer defines, a “video sequence” includes “[a] sequence of images” and “can be any video recording or media (e.g., an MPEG-1 file, a video tape, a video disc, a DVD, etc.).” EX1008 at 3:16-19. Accordingly, in my opinion, a POSITA would have found this limitation to be disclosed or at least rendered obvious.

d) Element 1/9[c]

184. In my opinion, this element was disclosed or at least would have been rendered obvious over Ozer in view of Zhao. As an initial matter, this element was disclosed or at rendered obvious in view of the teachings of Zhao, as described above for Ground I. The analysis for this element under Ground I is incorporated herein by reference.

185. In my opinion, Ozer further disclosed this limitation or at least rendered it obvious. Ozer explains that “[e]ach extracted region modeled with ellipses corresponds to a node in the graphical representation of the object of interest.” EX1008 at 7:9-11. The graph matching algorithm looks for “meaningful correspondence” where “[e]ach object part and meaningful combinations represent a class w where the combination of binary and unary features are represented by a feature vector X and computed off-line.” *Id.* at 7:11-19. As such, “[t]he computations needed for each node matching are then a function of the feature size and the previously matched nodes of the branch under consideration.” *Id.* at 7:25-

28. Accordingly, in my opinion, the step of measuring the object as a collection of components in each image was disclosed or at least rendered obvious to a POSITA.

e) Element 1/9[d]

186. In my opinion, this element was disclosed or at least would have been obvious over Ozer in view of Zhao. As explained previously, a POSITA would have looked to combine the teachings of Ozer with Zhao. Further, as described above for Ground I, this element was disclosed or at least rendered obvious in view of the teachings of Zhao. Accordingly, the analysis under Ground I is incorporated herein by reference and in my opinion this element was disclosed or at least rendered obvious. *See, supra*, Ground I, limitation 1/9(d).

f) Element 1/9[e]

187. In my opinion, this element was disclosed or at least would have been obvious over Ozer in view of Zhao. As explained previously, a POSITA would have looked to combine the teachings of Ozer with Zhao. Further, as described above for Ground I, this element was disclosed or at least rendered obvious in view of the teachings of Zhao. Accordingly, the analysis under Ground I is incorporated herein by reference and in my opinion this element was disclosed or at least would have been obvious. *See, supra*, Ground I limitation 1/9(e).

3. Claims 2/10

188. In my opinion, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 2/10. Claims 2 and 10 depend on Claims 1 and 9, respectively. As described above, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 2/10. *See, supra*, Ground I, Claims 2/10. The analysis for Ground I is incorporated herein by reference.

189. Further, in my opinion, Ozer disclosed or at least rendered obvious a spatio-temporal model for an object to be detected to be comprised of “detectors” for each of the collection of components. Ozer explains that each extracted part of an image “corresponds to a node in the graphical representation of the object of interest” (EX1008 at 7:9-11) and “[o]utput of the graph-matching algorithm is the classified object parts.” EX1008 at 7:34-35. Accordingly, Ozer disclosed or at least rendered obvious the claimed “detectors” where each of the collection of components are matched to nodes in a graphical representation of the object of interest. As such, in my opinion, Claims 2/10 was disclosed or at least would have been rendered obvious.

4. Claims 3/11

190. In my opinion, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 3/11. Claims 3 and 11 depend on Claims 1 and 9, respectively. As described above, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 3/11. *See, supra*, Ground I, Claims 3/11. The analysis for Ground I is incorporated herein by reference.

191. Further, the additional requirement of Claims 3/11 “wherein the spatio-temporal model is a graphical model comprising nodes corresponding to each of the collection of components and to the object” matches the Patent Owner’s purported construction of the term “spatio-temporal model.” Accordingly, in my opinion, Claims 3/11 were disclosed or at least rendered obvious for the same reasons discussed for limitation 1/9(a) and that purported construction. As described above, a POSITA would have found the purported spatio-temporal model disclosed or at least obvious in view of Zhao, as it describes a graphical model where “ribbons” serve as “nodes” that correspond to each of the collection of components (*i.e.*, the component body parts) and the object (*i.e.*, the merged parts that form a human). Similarly, a POSITA would have understood Ozer to teach the use of a graphical model comprising nodes [*i.e.*, the graphical

representation of the object of interest] corresponding to each of the collection of components [i.e., the classified object parts] and to the object [i.e., the grouped object parts, forming the object]. As such, in my opinion, Claims 3/11 was disclosed or at least would have been rendered obvious.

5. Claims 4/12

192. In my opinion, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 4/12. Claims 4 and 12 depend on Claims 1 and 9, respectively. As described above, Ozer in combination with Zhao disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference. Further, as explained above for Ground I, Zhao rendered obvious Claims 4/12. *See, supra*, Ground I, Claims 4/12. The analysis for Ground I is incorporated herein by reference.

193. In my opinion, a POSITA would have found Claims 4/12 disclosed or at least rendered obvious in further view of Ozer. Ozer teaches “Classifying Over Multiple Frames” where “[t]he movements of the object parts are described as a spatio-temporal sequence of feature vectors that consist of the direction of the object part movement. The system checks direction of the movements of the object parts for a number of frames and calculates the probabilities of the activities with the known activities by using Hidden Markov Models and chooses the pattern with the highest probability as the recognized activity in these frames (408).” EX1008 at

7:33-45. As such, a POSITA would have found it disclosed or at least obvious to detect an object according to measurements of the object as a collection of components determined from a “prior image” and a “later image” relative to the current image.

194. Notably, analyzing data in image frames from video by using measurements from adjacent frames, including previous and later frames, was widespread and commonly understood in image processing. *See* EX1011 (“Generating Spatiotemporal Models from Examples” which details an approach for generating spatio-temporal models that “allow[] accurate temporal extrapolation” and have “uses for tracking, recognition and data compression of deformable or articulated objects undergoing complex motions.”); *see also* EX1037 (describing a framework and analysis for Bayesian inference of visual shape and motion from video); EX1038 at Chapter 16.

195. Accordingly, in my opinion, Claims 4/12 were disclosed or at least rendered obvious over Ozer in combination with Zhao.

6. Claims 5/13

196. In my opinion, Ozer in combination with Zhao disclosed or at least rendered Claims 5/13 obvious. Claims 5 and 13 depend on Claims 1 and 9, respectively. As described above, Ozer in combination with Zhao disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully

herein by reference. Further, as explained above for Ground I, Zhao disclosed or at least rendered obvious Claims 5/13. *See, supra*, Ground I, Claims 5/13. The analysis for Ground I is incorporated herein by reference.

197. In my opinion, A POSITA would have found Claims 5/13 disclosed or at least rendered obvious in further view of Ozer. Ozer teaches “Classifying Over Multiple Frames” where “[t]he movements of the object parts are described as a spatio-temporal sequence of feature vectors that consist of the direction of the object part movement. The system checks direction of the movements of the object parts for a number of frames and calculates the probabilities of the activities with the known activities by using Hidden Markov Models and chooses the pattern with the highest probability as the recognized activity in these frames (408).” EX1008 at 7:33-45. By measuring the object parts for a “number of frames,” a POSITA would have known to define a “temporal window” in which measurements of components are passed to components detected in the current image. Indeed, for temporal inference problems defined on Markov chains, and temporal graphical models, well known probabilistic methods that include Gaussian sum filters (*see* EX1020) and use local linearizations to deterministically propagate Gaussian mixture models in space and time within a local temporal window of video image sequence frames. Similarly, particle filters (*see* EX1024) use Monte Carlo methods to stochastically update a set of weighted point samples of feature density

representations of probabilistic graphical models within a temporal window of video image sequences.

198. Accordingly, in my opinion, Claims 5/13 were disclosed or at least rendered obvious over Ozer in combination with Zhao.

7. Claims 6/14

199. In my opinion, Claims 6/14 were disclosed or at least would have been obvious over Ozer in view of Zhao. As explained previously, a POSITA would have looked to combine the teachings of Ozer with Zhao. Further, as described above for Ground I, Claims 6/14 were disclosed or at least would have been obvious in view of the teachings of Zhao. Accordingly, the analysis under Ground I is incorporated herein by reference and in my opinion Claims 6/14 were disclosed or at least would have been obvious here for the same reasons. *See, supra*, Ground I, Claims 6/14.

8. Claims 7/15

200. In my opinion, Claims 7/15 was disclosed or at least would have been obvious over Ozer in view of Zhao. As explained previously, a POSITA would have looked to combine the teachings of Ozer with Zhao. Further, as described above for Ground I, Claims 7/15 were disclosed or at least would have been obvious in view of the teachings of Zhao. Accordingly, the analysis under Ground

I is incorporated herein by reference and Claims 7/15 were disclosed or at least would have been obvious for the same reasons. *See, supra*, Ground I, Claims 7/15.

9. Claims 8/16

201. In my opinion, Claims 8/16 would have been obvious over Ozer in view of Zhao. As explained previously, a POSITA would have looked to combine the teachings of Ozer with Zhao. Further, as described above for Ground I, Claims 8/16 would have been obvious in view of the teachings of Zhao. Accordingly, the analysis under Ground I is incorporated herein by reference and Claims 8/16 would have been obvious for the same reasons. *See, supra*, Ground I, Claims 8/16.

D. Ground IV: Claims 1-16 Are Obvious Over TLP

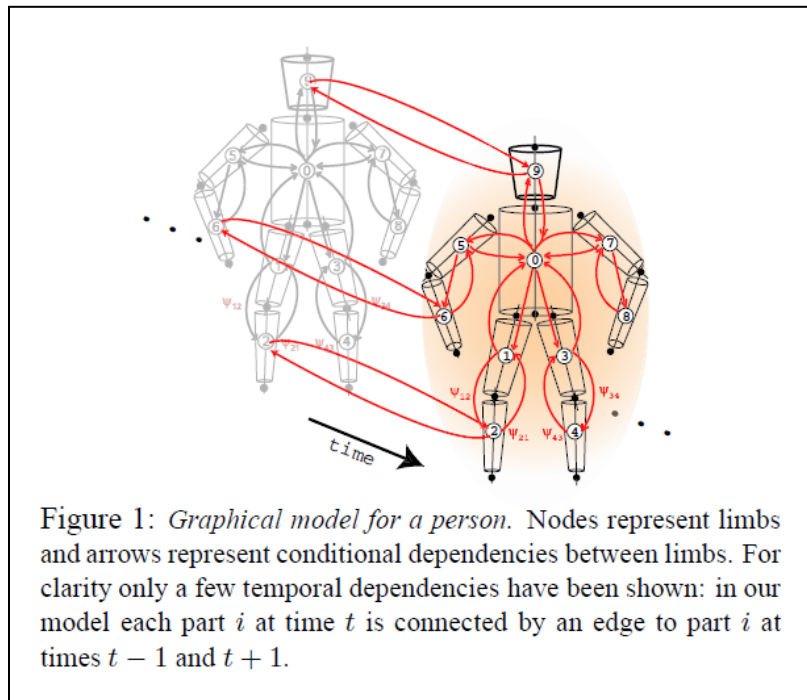
1. Claims 1/9

a) Element 1/9[preamble]

202. In my opinion, TLP disclosed or at least rendered this limitation obvious. Specifically, TLP describes “a probabilistic method for fully automatic 3D human detection and tracking” in the field of computer vision. EX1009 at 6. Accordingly, a computer implemented method for object detection would have been obvious to a POSITA. As a computer implemented method, a POSITA would have understood TLP to further include “a computer readable medium embodying a computer program” to perform that method.

b) Element 1/9[a]

203. In my opinion, TLP disclosed or at least rendered this limitation obvious. As part of its method for “human detection and tracking,” TLP provides a model for an object to be detected (*e.g.*, a human). TLP explains that “the body is represented by a graphical model in which each graph node corresponds to a body part (upper leg, torso, etc.). Each part has an associated configuration vector defining the part’s position and orientation in 3-space. ... Edges in the graphical model correspond to position and angle relationships between adjacent body parts in space and time, as illustrated in Figure 1.” EX1009 at 1.



EX1009 at 1. By incorporating spatial and temporal elements into a graphical model for an object to be detected, such as a person, TLP disclosed or at least rendered the claimed “spatio-temporal” obvious.

204. Further, even to the extent the term “spatio-temporal model” is construed as “a graphical model comprising nodes corresponding to each of the collection of components and to the object,” TLP still disclosed or at least rendered this element obvious. TLP’s model for an object to be detected is explicitly expressed as “a graphical model in which each graph node corresponds to a body part (upper leg, torso, etc.).” EX1009 at 2. Further, “[e]ach part has an associated configuration vector defining the part’s position and orientation in 3-space. Placing each part in a global coordinate frame enables the part detectors to operate independently while the full body is assembled by inference over the graphical model.” EX1009 at 2. A POSITA would have found the claimed spatio-temporal model disclosed or at least rendered obvious in view of TLP, as the nodes correspond to each of the collection of components and the object.

205. Accordingly, in my opinion, TLP disclosed or at least rendered this element obvious.

c) Element 1/9[b]

206. In my opinion, TLP disclosed or at least rendered this limitation obvious. TLP discloses a method for object detection and tracking for video. Accordingly, “providing a video comprising a plurality of images including the object” was disclosed or at least would have been obvious to a POSITA, because to detect an object in a video, providing the video itself would be required. EX1009 at

1 (“We illustrate the method by automatically tracking a walking person in video imagery using four calibrated cameras.”).

d) Element 1/9[c]

207. In my opinion, TLP disclosed or at least rendered this limitation obvious. TLP explains “our model of the body is a collection of loosely-connected limbs.” EX1009 at 1. To perform object detection, this body model is combined “with a probabilistic image likelihood model” and “[i]nferring the body pose in our framework is defined as estimating belief in the graphical model.” EX1009 at 4. “We define $\phi_i(X_i)$ to be the likelihood of observing the image measurements conditioned on the pose of limb i . Ideally this model would be robust to partial occlusions, the variability of image statistics across different input sequences, and variability among subjects. To that end, we combine a variety of cues including multi-scale edge and ridge filters following [20 [EX 1041]].” *Id.* at 4. Accordingly, measuring the object as a collection of components in each image was disclosed or at least would have been obvious to a POSITA, as the model itself is represented as “a collection of loosely-connected limbs” —i.e., a collection of components and a POSITA would have understood the collection of loosely-connected limbs to be equivalent to a collection of components. EX1009 at 1.

208. Notably, to describe the parts in the graphical model, “[e]ach part has an associated configuration vector defining the part’s position and orientation in 3-space,” as shown in Figure 2:

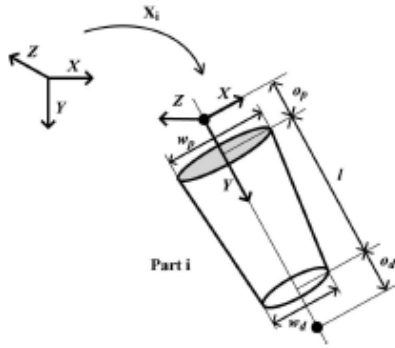


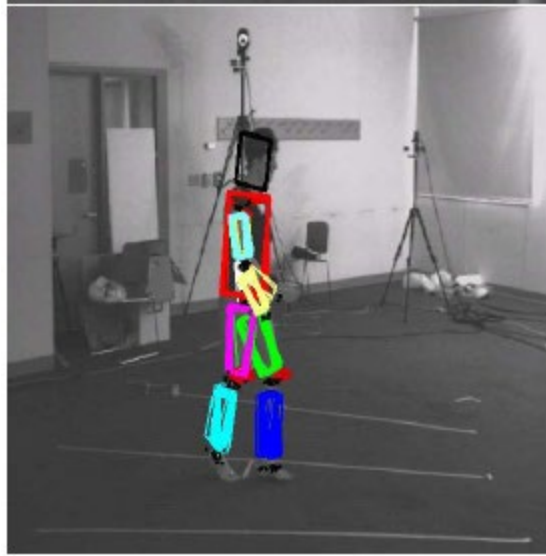
Figure 2: *Parameterization of part i.*

EX1009 at 2, 3.

209. “The estimated parameters $\mathbf{X}_i^T = (\mathbf{x}_i^T, \Theta_i^T)$ represent the configuration of the part i in a global coordinate frame where $\mathbf{x}_i \in \mathbb{R}^3$ and $\Theta_i \in \text{SO}(3)$ are the 3D position of the proximal joint and the angular orientation of the part respectively.”

EX1009 at 2-3. This matches the process described in the specification of the '980 Patent, which explains that nodes representing an object’s components “have an associated state vector $\mathbf{X}^T=(x,y,s)$ defining the component’s real-valued position and scale within an image.” EX1001 at 3:50-52.

210. Figure 6(c) illustrates, for example, the object being measured as a collection of components using the system defined in TLP:



c

EX1009 at 7, Figure 6(c) (illustrating “Samples from the belief distribution after 30 iterations of belief propagation”).

211. Accordingly, in my opinion, “measuring the object as a collection of components in each image” was disclosed or at least would have been obvious over TLP.

e) Element 1/9[d]

212. In my opinion, TLP disclosed or at least rendered this limitation obvious. TLP “present[s] a *probabilistic* method for fully automatic 3D human detection and tracking.” EX1009 at 6. Accordingly, the inference algorithm “combines the body model described above with a probabilistic image likelihood model” and defines “ $\phi_i(\mathbf{X}_i)$ to be the likelihood of observing the image measurements conditioned on the pose of limb i .” EX1009 at 4. To “be robust to

partial occlusions, the variability of image statistics across different input sequences, and variability among subjects,” TLP incorporates the use of a “Gibbs [likelihood] model”:

$$p(\mathbf{f} \mid \mathbf{X}_i) \propto \exp \left(- \sum_i \langle \lambda^{(i)}, \xi^{(i)}(\mathbf{f}, \mathbf{X}_i) \rangle \right)$$

EX1009 at 4.

213. Accordingly, a POSITA would have understood that the probabilistic method described is equivalent to determining a probability an object is in each image, and this limitation was disclosed or at least would have been obvious in view of TLP.

f) Element 1/9[e]

214. In my opinion, TLP disclosed or at least rendered this limitation obvious. As explained above for elements 1/9(d)-(e), TLP “present[s] a *probabilistic* method for fully automatic 3D human detection and tracking” and determines the likelihood (i.e., probability) that an object is in each image. Accordingly, a POSITA would have known to further compare the probabilities to a “threshold” to detect the object. Indeed, TLP describes the use of an “image likelihood model” and detecting the object is “defined as estimating belief in the graphical model.” EX1009 at 4. To infer the belief in a model, a POSITA would have known to compare a determined probability against a threshold. *See, e.g.,*

EX1038 at 15 (where “experimental evidence shows that specific choices of number of terms in the a priori densities and specific thresholds for the two approximations lead to a bound on the number of terms required to describe the a posteriori densities for any stage.”); EX1041 at 3 (“...the likelihood of observing the filter responses for an image is proportional to the ratio between the likelihood that the foreground image pixels are explained by the foreground object and the likelihood that they are explained by some general background” ... “This ratio is highest when the foreground (person) model projects to an image region that is unlikely to have been generated by some general scene but is well explained by the statistics of people.”).

215. Accordingly, a POSITA would have understood that the probabilistic method of detecting the object in any image is based on the comparison of the likelihood probabilities for each image to a threshold, and this limitation was disclosed or at least would have been rendered obvious in view of TLP.

2. Claims 2/10

216. In my opinion, TLP disclosed or at least rendered obvious Claims 2/10. Claims 2 and 10 depend on Claims 1 and 9, respectively. As described above, TLP renders Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference.

217. Further, as described above for limitation 1/9(a), TLP disclosed or at least rendered obvious a “spatio-temporal model” for an object to be detected. TLP further teaches the use of detectors for each of the collection of components. For example, according to TLP, “[w]e adopt simple limb and head detectors to provide ‘bottom-up’ information that is incorporated into the inference process at every time-step.” EX1009 at 1. Indeed, TLP acknowledges that “[m]any good head detectors exist and limb detectors have been used for some time.” EX1009 at 1 (citing prior art).

218. Accordingly, in my opinion, providing “detectors” for each of the collection of components of an object in an image and Claims 2/10 was disclosed or at least would have been obvious.

3. Claims 3/11

219. In my opinion, TLP disclosed or at least rendered Claims 3/11 obvious. Claims 3 and 11 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered obvious Claims 1/9. The analysis of Claims 1/9 is incorporated fully herein by reference.

220. Further, the additional requirement of Claim 3/11 “wherein the spatio-temporal model is a graphical model comprising nodes corresponding to each of the collection of components and to the object” matches the Patent Owner’s purported construction of the term “spatio-temporal model.” Accordingly, Claim

3/11 was disclosed or at least would have been obvious for the same reasons discussed for limitation 1/9(a) and that purported construction.

4. Claims 4/12

221. In my opinion, TLP disclosed or at least rendered Claims 4/12 obvious. Claims 4 and 12 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 4/12 is incorporated fully herein by reference.

222. As described for elements 1/9(d)-(e) above, TLP presents a probabilistic method that determines the probability that an object is in each image based on identified body parts, which includes temporal dependencies. As TLP explains, “in our model, each part i at time t is connected by an edge to part i at times $t - 1$ and $t + 1$.” EX1009 at 1. Accordingly, TLP uses “probabilistic models for the temporal evolution of each limb (forward and backward in time). Human pose and motion estimation is then solved with non-parametric belief propagation using a variation of particle filtering that can be applied over a general loopy graph.” EX1009 at 1. Accordingly, in my opinion, TLP disclosed or at least rendered obvious the step of detecting an image according to measurements of the object as a collection of components determined from a prior image (i.e., $t - 1$) and a later image (i.e., $t + 1$), relative to the current image (i.e. t).

5. Claims 5/13

223. In my opinion, TLP disclosed or at least rendered Claims 5/13 obvious. Claims 5 and 13 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference.

224. As described above for limitation 1/9(a), TLP describes a spatio-temporal model for an object to be detected. As TLP explains, “in our model, each part i at time t is connected by an edge to part i at times $t - 1$ and $t + 1$.” EX1009 at 1. Accordingly, TLP uses “probabilistic models for the temporal evolution of each limb (forward and backward in time). Human pose and motion estimation is then solved with non-parametric belief propagation using a variation of particle filtering that can be applied over a general loopy graph.” EX1009 at 1. TLP further discloses using “a **3-frame windowed** smoothing algorithm for our tracking results where the estimates at time t are based on observations at time $(t - 1, t + 1)$.” EX1009 at 5. Accordingly, a POSITA would have understood that TLP describes providing a “temporal window” defining a “plurality of images” in which measurements of components detected are passed to components detected in the current image.

6. Claims 6/14

225. In my opinion, TLP disclosed or at least rendered Claims 6/14 obvious. Claims 6 and 14 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference.

226. As described above, TLP describes determining the probability that an object is in each image. *See* limitation 1/9(d). As part of that process, TLP explains “the body is represented by a graphical model in which each graph node corresponds to a body part (upper leg, torso, etc.). Each part has an associated configuration vector defining the part’s position and orientation in 3-space.” EX1009 at 2. By determining the part’s position and orientation in 3-space, a POSITA would have known to determine the probability for a “position” of the object in each image.

227. Further, “[e]ach part/limb is modeled by a tapered cylinder having 5 fixed and 6 estimated parameters,” which include “the part length, width at the proximal and distal ends and the offset of the proximal and distal joints along the axis of the limb...” EX1009 at 2. By determining the part’s parameters, such as length and width, a POSITA would have known to determine the probability for a “size” of the object in each image. Accordingly, in my opinion, a POSITA would have recognized that TLP disclosed or at least rendered obvious the step of

“determining the probability that the object is in each image comprises determining the probability for a position and a size of the object in each image”.

7. Claims 7/15

228. In my opinion, TLP disclosed or at least rendered Claims 7/15 obvious. Claims 7 and 15 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference.

229. As described above for limitation 1/9(e), a POSITA would have known to compare a determined probability against a threshold. It would have been further known that this threshold is determined empirically, as opposed to arbitrarily chosen. Indeed, TLP explains that “[c]onditional probabilities relating the 3D pose of connected limbs are learned from motion captured training data [i.e., empirically].” EX1009 at 1. As a POSITA would have understood, determining a threshold “empirically” is done via observation or experimentation. *See, e.g.*, EX1017; EX1018; EX1040 at 4 (explicitly disclosing that the probabilistic thresholds are experimentally set using techniques such as ADABOOST: “The key insight is that smaller, and therefore more efficient, boosted classifiers can be constructed which reject many of the negative sub-windows while detecting almost all positive instances (i.e., the threshold of a boosted classifier can be adjusted so that the false negative rate is close to zero). Stages in the cascade are

constructed by training classifiers using AdaBoost and then adjusting the threshold to minimize false negatives.”). By learning the conditional probabilities through training data, a POSITA would have known to determine a threshold “empirically.” Accordingly, in my opinion, Claims 7/15 was disclosed or at least would have been obvious.

8. Claims 8/16

230. In my opinion, TLP rendered Claims 8/16 obvious. Claims 8 and 16 depend on Claims 1 and 9, respectively. As described above, TLP disclosed or at least rendered Claims 1/9 obvious. The analysis of Claims 1/9 is incorporated fully herein by reference.

231. As described for Claim 1/9, TLP presents a probabilistic method for object detection. *See* limitations 1/9(d)-(e). The joint probability distribution formulated in Claims 8/16 would have been well known to a POSITA and is the very same pairwise Markov field representation of a spatio-temporal graphical model, which is cited in TLP. *See* EX1009 at 4 (“we use a form of non-parametric belief propagation [13 [EX1019 PAMPAS], 25 [EX1012 Sudderth]].”). The equation of Claim 8/16 represents a typical factorized form of the joint probability distribution $P(\dots)$ for a graphical model G of an object O with state X_i^O at time frame i of the video and $\{k = 1..N\}$ components state X_i^{C-k} . This form of the joint probability would have been well known to a POSITA as a pairwise Markov field

representation of a two-level spatio-temporal graphical model. The first level representing the object state at each time frame, and the second level for individual components states, and captured by their pairwise probabilistic relationships as defined by edges in the graphical model *Id.* (citing EX1012 (Sudderth) at 2, Sec 2.1).

X. SECONDARY CONSIDERATIONS

232. I am unaware of any evidence of secondary considerations that would support a finding of non-obviousness. In my opinion, the asserted prior art demonstrates there is no evidence of failure by others and that the features recited in the Challenged Claims were readily available in the prior art. I also believe there is a strong case of obviousness, as described in the analysis above. Should Patent Owner present any evidence regarding secondary considerations, I reserve the right to respond.

XI. CONCLUSION

233. For the foregoing reasons, in my opinion, a POSITA would have found Claims 1-16 to be obvious in view of the prior art, as detailed in Grounds I-IV.

Dated: May 24, 2023

Respectfully submitted,



Chandrajit Bajaj, Ph.D.