

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

IPR2023-00924

**SUPPLEMENTAL DECLARATION OF CHANDRAJIT BAJAJ, PH.D. IN
SUPPORT OF PETITIONERS' REPLY**

TABLE OF CONTENTS

I.	Introduction.....	1
II.	Dr. Saber’s Claim Construction of Element 1[e]/9[e] Is Inconsistent with the ’980 Specification and File History.....	3
III.	The ’980 Patent Does Not Distinguish the 1-Frame Windowed Smoothing Algorithm Scenario from the Purported Invention.....	13
IV.	A POSITA Would Have Known the Benefits of Extrapolating Motion Information from Prior and Later Frames	21
V.	Steffens’ Wavelet-Based Approach Is Not Fundamentally Different from or Incompatible with Zhao’s Contour-Based Approach	23
VI.	A POSITA Would Have Understood that Ozer’s “Activity” Classification Encompasses Object Tracking	40
VII.	The Joint Probability Distribution Recited in Claim 8/16 Is the Same Pairwise Markov Field Representation of a Spatio-Temporal Graphical Model Disclosed by TLP and Prior Art Cited in My First Declaration.....	44
VIII.	Conclusion	60

TABLE 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 1539–46 (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, 702–18, 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 613–20 (2003)
1020	D. L. Alspach & H. W. Sorenson. <i>Nonlinear Bayesian estimation using Gaussian sum approximations</i> , IEEE Trans. AC, 17(4): at 439–48 (August 1972)
1021	James M. Coughlan & Sabino J. Ferreira, <i>Finding deformable shapes using loopy belief propagation</i> , ECCV 453–68 (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> , Int'l. Conf. on Computer Vision at 94–101 (1999)
1026	Y. Wu, G. Hua & T. Yu, <i>Tracking articulated body by dynamic Markov network</i> , Int'l. Conf. on Computer Vision at 1094–101 (2003)
1027	Christopher Bregler & Jitendra Malik, <i>Tracking people with twists and exponential maps</i> , CVPR, at 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, at 669–76 (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 447–54 (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)
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)

Exhibit No.	Description
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 at 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> , 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)

Exhibit No.	Description
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)
1044	Patent Assignment for U.S. Patent No. 7,436,980
1045	Patent Assignment for U.S. Patent No. 7,436,980
1046	Application Data Sheet 37 C.F.R. 1.76 for U.S. Patent No. 9,189,886
1047	Patent Assignment for U.S. Patent No. 7,436,980
1048	<i>CVPR Technical Program: Call for Papers</i> , IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, June 27 – July 2, 2004
1049	ResearchGate Web Page for TLP
1050	Declaration of Dr. Sylvia Hall-Ellis relating to TLP
1051	Deposition Transcript of Eli Saber, Ph.D. (May 30, 2024)
1052	Supplemental Declaration of Dr. Chandrajit Bajaj
1053	U.S. Patent No. 6,597,801 to Cham <i>et al.</i>

** Unless otherwise indicated, all emphasis added.

I, Chandrajit Bajaj, declare as follows:

I. INTRODUCTION

1. I am over the age of eighteen (18) and otherwise competent to make this Supplemental Declaration. I previously submitted a Declaration (“First Declaration,” Exhibit 1004) containing my opinions in this proceeding.

2. I am making this Supplemental Declaration at the request of Petitioners Meta Platforms, Inc.; Instagram, Inc.; WhatsApp LLC; Meta Platforms Technologies, LLC; and GIPHY, Inc. (“Petitioners”) in the *Inter Partes* Review of U.S. Patent No. 7,436,980 (“’980 Patent”) about certain issues that I understand from counsel relate to arguments presented in Patent Owner’s Response (“POR,” Paper 26) and the accompanying Declaration of Dr. Eli Saber (“Saber Declaration,” Exhibit 2004) filed on March 13, 2024, as well as the deposition testimony of Dr. Saber (“Saber Transcript,” Exhibit 1051) provided on May 30, 2024.

3. 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.

4. I was not asked by Petitioners to provide my opinion about every argument, contention, opinion, or piece of evidence in Patent Owner’s Response,

the Saber Declaration, or the Saber Transcript. Rather, I only provide my opinion regarding certain issues, as explained in more detail below. Accordingly, I have not formed an opinion about any of the arguments, contentions, opinions, or pieces of evidence in Patent Owner's Response, the Saber Declaration, or Saber Transcript except as I address below (and except to the extent a relevant opinion may be found in my First Declaration, which I offered before Patent Owner's Response, the Saber Declaration, or the Saber Transcript).

5. In the preparation of this Supplemental Declaration and forming my opinions, I have considered:

- the Board's Institution Decision ("ID"), Paper 11;
- the documents identified in the Table of Exhibits table above;
- the relevant legal standards, including the standard for obviousness, as I understand from counsel;
- any additional authoritative documents as cited in the body of this Supplemental Declaration; and
- my own knowledge and experience based upon my work in the areas of computer science, applied mathematics, image processing, and computer vision and visualization as described below.

II. DR. SABER’S CLAIM CONSTRUCTION OF ELEMENT 1[E]/9[E] IS INCONSISTENT WITH THE ’980 SPECIFICATION AND FILE HISTORY

6. Dr. Saber set forth a claim construction for claim element 1[e]/9[e].

Ex2004, Section VII. That term recites “detecting the object in any image upon comparing the probabilities for each image to a threshold for detecting the object.” I understand that Dr. Saber disagrees with the Board’s provisional determination as to element 1[e]/9[e]. ID at 27. I understand the Board disagreed that “comparing multiple probabilities from multiple images to a threshold is inconsistent with the specification’s disclosure of this term.” ID at 27.

7. Dr. Saber’s opinion is purportedly based on “the surrounding claim language, the patent’s specification, and the patent’s file history.” Ex2004, Section VIII. I disagree with Dr. Saber that element 1[e]/9[e] requires comparing multiple probabilities from multiple images to a threshold for detecting the object in any given image. I believe Dr. Saber is reading in limitations from the ’980 specification, which a POSITA would not understand to be a requirement of claims 1 and 9. For example, when asked during his deposition whether claims 1 and 9 require the spatio-temporal model to be graphical (Ex1051 at 69:3-21), Dr. Saber responded:

3 Q But does the claim require the model to be
4 graphical?
5 A Well, when you read the claim in light of
6 the specification, as a POSITA I'm reading the
7 claim, you have an object as a collection of
8 components, so you have components that make up
9 the object, how you're putting these components
10 together in light of the specification, they
11 propose this graphical, two-layer model.
12 Q Does the claim require a two-layer model?
13 ATTORNEY EKLEM: Objection to form.
14 Q Claim 1.
15 ATTORNEY EKLEM: Object to the form.
16 A Again, when you read the claim in light of
17 the specification, a POSITA would understand that
18 you are formulating the object, measuring the
19 object as a collection of components in each
20 image. So you have that essentially design that's
21 being proposed in this patent.

8. The language of claims 1 and 9 and the '980 specification does not support Dr. Saber's understanding and interpretation. For example, as noted in the Ex2004 at 25, claims 1 and 9 perform the steps of providing, measuring, determining, detecting, and comparing without using or referring to a graphical two-layer model.

9. Dr. Saber's understanding of the language of claim element 1[d]/9[d] is also influenced by those (unspecified) graphical two-layer models from which he states the provided spatiotemporal model is created (Ex1051 at 122:2-17).

2 The next step is, you're determining the
3 probability that the object is in each image.
4 That probability is based on the spatiotemporal
5 model that you have created in those graphical
6 two-layer models. It's not just based on one
7 component.
8 So you have a spatiotemporal model, you
9 have a two-layer graphical model. You're going to
10 use that to determine the probability that the
11 object is in each image, because that's the only
12 thing would make sense to a POSITA.
13 And then you detect the object in any
14 image of those video, collection of images,
15 plurality of images, based on comparing the
16 probabilities for each image to a threshold for
17 detecting the object.

10. Dr. Saber's understanding of claim element 1[e]/9[e] is also influenced by the specific spatio-temporal model that is created from the graphical two-layer model, despite any mention of this in the claim language. *See e.g.*, Ex1051 at 118:19-119:3 ("So that sentence, 'detecting the object in any image instance {sic},' is essentially using the spatiotemporal model that you have created,

and you are doing that based on comparing the probabilities for each image that you've computed from this spatiotemporal model to a threshold for detecting the object").

11. When questioned during his deposition (Ex1051 at 69:22-71:11) whether the '980 specification contemplated a single object layer, Dr. Saber acknowledged the "alternative potential," but concluded that the '980 Patent is proposing "justification why a two-layer model would be more beneficial":

5 A Yeah, so that's an alternative potential.
6 But if you read before that, you can see, "To
7 integrate temporal constraints, the spatial
8 graphical models are extended over time to an
9 arbitrary length temporal window."

10 So you have a spatial graphical model and
11 you're doing a temporal thing, as you see in
12 Figure 2 of the patent.

13 "The resulting spatiotemporal graphical
14 models are shown in Figure 2A and 2B." And if you
15 look at those figures, you'll see essentially a
16 time axes on top, and with time going from left to
17 right, and then you see the object layer and the
18 components layer, and essentially illustrating a
19 graphical model that's temporally built.

20 "Having a two-layer graphical model allows
21 the inference process to reason explicitly about
22 the object as a whole" - e.g., two or three,

1	because you can talk about the object by itself -
2	"and reduce the complexity of the graphical model
3	by allowing the assumption of the conditional
4	independence of components." Because then you can
5	get into the components part of the object. So
6	there is an advantage for doing the two-layer
7	model as they're proposing.
8	Alternatively, he says you can have a
9	single layer model. But obviously what they're
10	proposing is they're giving you the justification
11	why a two-layer model would be more beneficial.

12. Elements 1[c]/9[c], 1[d]/9[d] and 1[e]/9[e] recite:

1[c]/9[c]: measuring the object **as a collection of components in each image**;

1[d]/9[d]: determining **a probability** that the object is in **each image**; and

1[e]/9[e]: detecting the object in any image upon comparing **the probabilities for each image** to a threshold for detecting the object.

13. These claim 1 and 9 limitations clearly specify image processing steps, and on an image frame by image frame basis, consisting of measuring the **object as a collection of components**, determining a probability that the object is present, and detecting the (presence of the) object in any image by comparing the (collection of component) probabilities for each image to a threshold. The prior limitations recite:

1[a]/9[a]: providing a spatio-temporal model for an object to be detected;

1[b]/9[b]: providing a video comprising a plurality of images including the object;

14. As a precursor to image by image processing steps, limitation

1[a]/9[a] provides a spatio-temporal model for an object, so each of the limitations

1[c]/9[c], 1[d]/9[d] and 1[e]/9[e] can be performed for each of the images

comprising the video, provided by limitation 1[b]/9[b]. A POSITA would have

understood that claims 1 and 9 are not limited to a graphical model, nor a two-layer

graphical model, but a spatio-temporal model that consists of a set or collection of

components, each with spatial dimensions that are measurable, and with the ability

of this component collection model to change configuration in time.

15. Despite all evidence to the contrary, and including the prior art that I cited in my First Declaration, Dr. Saber testified (Ex1051 at 60:11-19) that a part of the '980 Patent's inventive contribution is providing a spatio-temporal model to conduct object detection:

11	Q So the use of a spatiotemporal model to
12	detect an object, was that well-known in the 2004
13	time frame?
14	ATTORNEY EKLEM: Objection to form.
15	A Again, I think you're misunderstanding
16	what I'm saying, perhaps. So he's -- in the '980
17	patent, he's providing -- the authors are
18	providing a spatiotemporal model to do this
19	detection. It's part of their contribution.

16. In light of the '980 specification, a POSITA would have understood that “a probability” that the object recited in 1[d]/9[d] is in each image is determined as a function of the respective probabilities of the “collection of components” recited in 1[c]/9[c]. The '980 specification indicates by way of example that a joint probability distribution for the spatio-temporal model with N components is trained to determine whether an object is present in an image. *See* '980 Patent at 3:53-65 and 5:25-67. In order to detect the object in “any” one of the plurality images as recited by 1[e]/9[e], the '980 Patent indicates that AdaBoost can be used to boost the confidence that the “collection of components” are present in the image.

17. According to the '980 specification (at 5:32), boosted detectors are trained for each component N using AdaBoost. Each component N has an associated weak classifier and the AdaBoost procedure learns a weighted

combination of the base weak classifiers. The output of the AdaBoost classifier is a confidence that the object is present in an image where an object is considered present if the determined confidence is greater than or equal to the threshold recited at 5:65 of the '980 Patent. The confidence is converted to a likelihood function that the object is detected in each image. *See* '980 Patent at 6:1-3. Thus, a POSITA would have understood that “a probability” that the object recited in 1[d]/9[d] is in each image is determined as a function of the respective probabilities of the “collection of components” recited in 1[c]/9[c].

18. Dr. Saber indicates that his interpretation of 1[e]/9[e] is based in part on the '980 file history. Ex2004 at ¶63. However, I do not see any discussion of the '980 file history in his declaration in support of his interpretation of 1[e]/9[e]. I have reviewed the '980 file history (Ex1002) and believe the discussion of the “close” prior art discussed by the examiner in the Notice of Allowance (p. 163) reflects that the examiner did not understand the claim limitation 1[e]/9[e] to require comparing multiple probabilities from multiple images to a threshold.

19. The examiner acknowledged (Ex1002 at 163) 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).

20. As I mentioned in my First Declaration, Ex1004 at ¶44, while the examiner found (in error) that Cham (Ex1053) did not describe a spatio-temporal model, he correctly concluded that Cham disclosed the probabilistic framework of the '980 Patent claims. The examiner's conclusion that Cham teaches "comparing a total probability with a predetermined threshold" reflects to me that he too understood limitation 1[e]/9[e] to require comparing the probabilities of the object features to a threshold on an image frame by image frame basis in order to determine the probability that the object is present.

21. The examiner's conclusion is consistent with my understanding of Cham. Similar to the '980 Patent, Cham uses a probabilistic framework to identify occurrences of an object model in an image (6:1-7):

1) an "object model" which describes a target to be located in **an input image**; 2) a **"comparison function"** that determines how well a specific instantiation of the object model matches the image data; and 3) a "search strategy" that specifies how to systematically identify occurrences of the object model in **the image**. {Emphasis added.}

A POSITA would have understood from Cham that the object model is decomposed into two main components including (i) a set of features and (ii) a vector of parameters "that encodes the configuration of the features in the image."

Ex1053, 6:8-13. And a POSITA would have further understood from Cham that the search for the object could be based on maximizing a "comparison function."

Ex1053, 3:48-50. According to Cham (3:58-66):

The number of required search operations is determined by first finding search regions within a feature space, where each region has an associated probability density which exceeds some predetermined threshold. A total probability is then formed by summing the probabilities associated with each of the found search regions. If the total probability is less than the predetermined probability, the threshold is lowered and new regions are found and the probabilities summed.

22. A POSITA would have understood this passage, among others, to mean that the total probability that an object is in an image frame is compared to a “predetermined probability” threshold on an image frame-by-image frame basis. The threshold is empirically adjusted; however, the probabilities are always determined for a single frame. A POSITA, and apparently the ’980 Examiner, would *not* have understood Cham to teach comparing *multiple probabilities* from *multiple images* to a threshold to detect the object.

III. THE ’980 PATENT DOES NOT DISTINGUISH THE 1-FRAME WINDOWED SMOOTHING ALGORITHM SCENARIO FROM THE PURPORTED INVENTION

23. Dr. Saber concludes in ¶69 of his declaration, Ex2004, that “interpreting the claims to encompass detecting an object based on a probability for a single image (*i.e.*, the $w=1$ ‘single frame component-based fusion’) would ignore invention’s purpose and function described in the specification.” I disagree.

24. In support of his conclusion, Dr. Saber opines that (i) the 1-frame windowed smoothing algorithm (“ $w=1$ ”) does not achieve the purpose or function of the ’980 “invention” and (ii) the ’980 Patent demonstrates via a real-world

comparison of its 3-frame ($w=3$) analysis with the $w=1$ scenario that the $w=1$ scenario had “poor” performance. I address both (i) and (ii) next.

25. According to Dr. Saber (§67):

A POSITA would have also considered that object detection based on a multi-image analysis is fundamental to the purpose of the invention, which is to provide “[o]bject detection that considers the consistency of the detection **over time**,” meaning based on multiple images. {Emphasis in original.}

26. Further, during his deposition (Ex1051 at 129:13-131:12), Dr. Saber testified that the '980 Patent “invention” **requires** a multi-image analysis for initialization of the spatio-temporal model:

13	Q Everything you just stated in your answer
14	is applicable to a single frame, W equals 1. It
15	has nothing to do with each frame having a
16	separate probability, and comparing each one of
17	those probabilities to a threshold, correct?
18	ATTORNEY EKLEM: Objection to form.
19	A Totally incorrect. Absolutely incorrect.
20	Q Well, let's do it piecemeal. What did you
21	just say that's inapplicable to a single frame?
22	ATTORNEY EKLEM: Objection to form.

1 A Well, you have a graphical model that's
2 spatiotemporal, right? So you have defined the
3 graphical model if you've defined the
4 spatiotemporal apportionment. That by itself,
5 that temporal, is more than one frame.

6 Q How about the initialization?

7 A Well, the initialization, you initialize
8 it on the first, let's say two or three frames
9 that you've receive, but you still need the
10 temporal aspect of it.

11 Q For the initialization you need the
12 temporal aspect?

13 A Even with the initialization. So you're
14 going to get the first, say, series of frames, you
15 get first four, five, six; you initialize the
16 spatiotemporal model. You need to collect --
17 buffer a few frames to do the spatiotemporal
18 model.

19 Q So the initialization can't be based on
20 pattern recognition and scaling on the first
21 image?

22 ATTORNEY EKLEM: Objection to form.

```
1      A  You would need to collect -- because you
2      have to initialize the spatiotemporal model.  You
3      have a model that has spatial parameters and
4      probabilities and temporal parameters and
5      probabilities.  So you can't just say, well, I'll
6      just get my image, one image, and I'll do this.
7      You're going to collect a series of images 10, 15
8      frames, whatever.  You're going to collect a
9      series of images.  You're going to initialize your
10     model in those images, and then start basically
11     moving it as you add more images into -- into your
12     sequence.
```

I disagree.

27. A POSITA would have understood from the '980 claims and specification that providing a spatio-temporal model (*e.g.*, human or car) consists of specifying the object to be detected and tracked, as a set or collection of components, each with spatial dimensions that are measurable, and with the ability of this component collection object model to change configuration in time. The **initialization** is the beginning of the sequence of detection and tracking of the object presence in the first image (for *e.g.*, of a human walking sequence or a moving car sequence) and follows the image processing limitations of claims 1/9[c]-[e] of measuring the object as a collection of components in the first image, determining a joint probability that the object is present in the first image, and

further detecting the (presence of the) object in the first image by comparing the (collection of component) probabilities of the first image to a threshold. After the initialization, the process is repeated for each image on an image frame by image frame basis of the provided video image sequence.

28. Furthermore, when determining the joint probability (likelihood) of an object and each of its components being present to calculate the joint probability (likelihood) that an object is present in a current image frame $I(j)$ could be accomplished **without** fusing or combining probabilities from a previous image frame $I(j-1)$ and/or future image frame $I(j+1)$ via Belief Propagation (“BP”).

Note, that in general the operation is specified as working on previous image frame $I(j-(w-1)/2)$ and/or future image frame $I(j+(w-1)/2)$ for $w \geq 1$. The process for $w=1$ is also included and accomplishes all its image processing steps as described by operating only on the image frame $I(j)$, as the previous and future in this scenario is only the current $I(j-0) = I(j)$.

29. In this single image frame processing without windowing (*i.e.*, the $w=1$ scenario), BP is performed between the likelihoods of multiple components present within the image frame for their spatial location and spatial connectivity of the multi-component spatio-temporal model. That is, even when $w=1$, BP is still applicable to the detection and determination of the likelihood probabilities of the multiple components of the model in each image frame. Thus, even though the

ability to disambiguate an object in a current frame by referencing the probabilities in the previous or future frame may be lost when $w=1$, the ability to disambiguate the object in the current frame alone is still possible after multiple iterations of the BP algorithm. *See, e.g.,* Ex1019 (“After initialisation the PAMPAS algorithm using the Gibbs sampler ... for the circle’s messages and belief is iterated without further recourse to the “feature detector” information.... After two or three iterations the belief distribution has converged on the object.”)

30. Dr. Saber did acknowledge during his deposition (Ex1051 at 94:9-19), however, that the first time an object is detected in an image frame, an object must be detected without the benefit of motion information from a prior image frame:

9	A So you detect it. You don't have a
10	velocity at that point because you only have one
11	frame. And let's say you're capturing that 30
12	frames a second. So I just have a first frame.
13	Now I see it in the next frame, I start to
14	establish, okay, it looks like it's moving at
15	60 miles an hour. So I start to establish a
16	rapport, right? And as you see it more and more
17	frames, your models become more and more accurate
18	in tracking the object because you're -- you have
19	historical information from prior frames.

31. The '980 Patent (at 7:56-8:3) compares a “simple fusion scheme” to a 3-frame spatio-temporal object model, which was shifted in a tracking centric way

over time, to disambiguate the BP inference process for determining the probability that an object is in each image. Dr. Saber opines in ¶69 of his Declaration that the “simple fusion scheme” used in the experimental comparison is the $w=1$ scenario and that “A POSITA would have understood the specification to be contrasting the claimed invention with a known technique to highlight the function of the claimed invention.” I disagree. To the extent that a POSITA would have understood the “simple fusion scheme” used in the experimental comparison to be the $w=1$ scenario, the POSITA would have understood the ’980 specification to be highlighting the benefits of using the 3-frame smoothing window, not distinguishing the $w=1$ scenario from the subject matter recited by the ’980 claims.

32. While employing a 3-frame smoothing window may improve detection and tracking relative to the $w=1$ scenario, the ’980 specification expressly indicates that w is an odd integer ≥ 1 . If $w=1$ was not intended to be included, then the specification could have indicated that w is an odd integer > 1 or w is an odd integer ≥ 3 . That is not the case here.

33. Further (and contrary to Dr. Saber’s deposition testimony at, for example, 130:11-131:12 and 225:18-226:13), the $w=1$ smoothing window scenario is completely applicable to the image processing limitations (steps) 1[c,d,e]/9[c,d,e] and to the providing steps 1[a,b]/9[a,b].

34. The image frame by image frame processing limitations process each image frame individually without any fusion from neighboring frames. Namely, element 1[c]/9[c] “measuring the object as a collection of components in each image” is performed on the current image, relying only on the image processing in that current frame. A POSITA would have understood that this measurement image processing element 1[c]/9[c] is individually performed on a current image frame, and sequentially image by image for each image of a video.

35. With regard to elements 1[d]/9[d] “determining a probability that the object is in each image” and 1[e]/9[e] “detecting the object in any image upon comparing the probabilities for each image to a threshold for detecting the object,” the image processing is similar and accomplished individually for a current image frame, and sequentially for all image frames of the video, an image frame at a time. That is, under the $w=1$ scenario, steps 1/9[c], [d] and [e] is performed image frame by image frame of the video image sequence, without any fusion from previous or future image frames, using the multi-component object probabilities for detecting the presence of the object in each image frame. A POSITA would have known that windowed smoothing using probabilities from previous and future frames was optional and straightforward. *See Attractive people*, Ex1015 at 7 (“forward-backward smoothing, either over a time window or an entire sequence, is straight forward.”).

IV. A POSITA WOULD HAVE KNOWN THE BENEFITS OF EXTRAPOLATING MOTION INFORMATION FROM PRIOR AND LATER FRAMES AS CLAIMED IN CLAIMS 4/12

36. Dr. Saber indicates in ¶119 (regarding dependent claims 4/12) that I provide no explanation how the references discussed in my First Declaration support my conclusion that 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.

37. As I stated in my First Declaration (¶109), 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 Attractive People*, Ex1015 at 7 (“forward-backward smoothing, either over a time window or an entire sequence, is straight forward.”).

38. A POSITA would have further understood the spatio-temporal inference of a correct walking sequence would operate on each image frame for detecting spatial information of the human, and additionally that tracking occurs across multiple consecutive image frames, spanning each image frame, so the entire spatio-temporal walking sequence is a correct spatio-temporal detection in each image frame and for the entire walking sequence. *See, e.g.*, Ex1011 at 1, 9 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.”

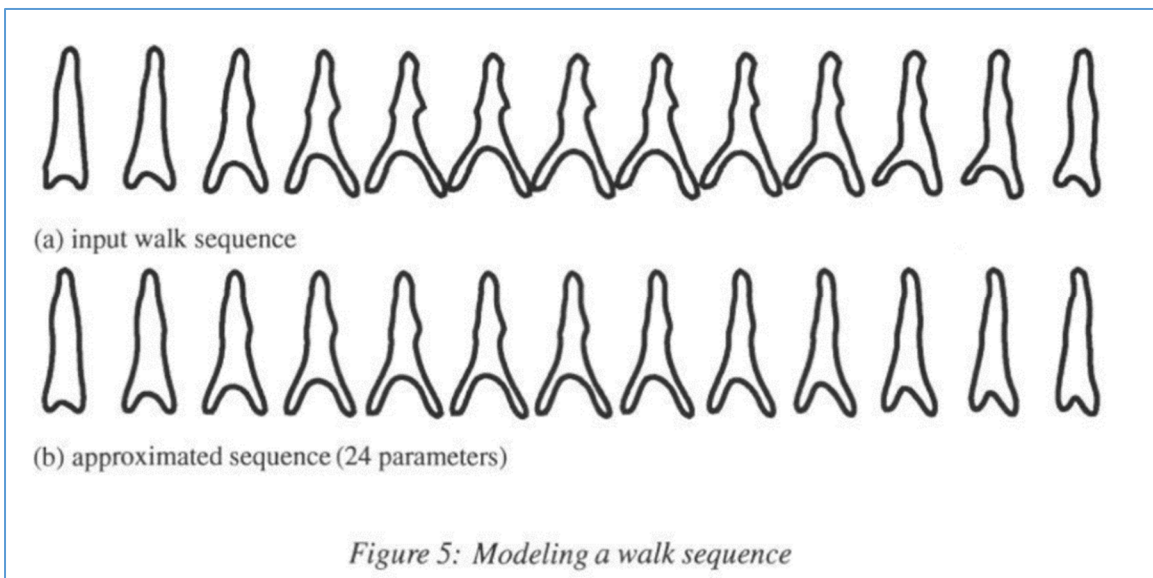
See also Ex1011 at 3 below:

Hence we assume an observed set of matched, aligned shape vectors consisting of nodal (or point) positions observed over short intervals of time. e.g. a set of shape vectors $\mathbf{x}^{(k)}$ each consisting of n control points.

$$\mathbf{x}^{(0)} = (P_x^1, P_y^1, \dots, P_x^n, P_y^n)$$

with $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ observations of the nodes at time $t = 0, \Delta t, 2\Delta t$. From this data set a set of nodal displacements $\mathbf{u}^{(k)}$ is extracted by subtracting off the mean shape vector. The corresponding nodal velocities $\dot{\mathbf{u}}^{(k)}$ and nodal accelerations $\ddot{\mathbf{u}}^{(k)}$ are then calculated by finite difference approximations.

39. Figure 5 of Ex1011 shows the approximation of a spatio-temporal shape model walking sequence (bottom row (b)) $\mathbf{x}^{(k)}$, given the input video image sequence (top row (a)), through measurement estimations of shape displacements **in each image frame (k)**, and finite difference estimations of velocity $\dot{\mathbf{u}}^{(k)}$, acceleration $\ddot{\mathbf{u}}^{(k)}$ between 3 adjacent image frames $t = 0, \Delta t, 2\Delta t$, and with temporal consistency along the entire walking sequence.



V. STEFFENS' WAVELET-BASED APPROACH IS NOT FUNDAMENTALLY DIFFERENT FROM OR INCOMPATIBLE WITH ZHAO'S CONTOUR-BASED APPROACH

40. Dr. Saber states in ¶129 of his declaration that “[a] POSITA would have understood that Steffens’ wavelet-based approach is fundamentally different from, and thus incompatible, with Zhao’s contour-based approach.” This not true.

41. A POSITA would have known that Steffens and Zhao each similarly concern methods for spatio-temporal **object detection and tracking in video images** based on multi-component dynamic object models and similarity matching in image frames incorporating motion information across frames. During his deposition (Ex1051 at 258:9-259:14), Dr. Saber’s answer to a direct question of whether Steffens employs a spatio-temporal model, instead acknowledges that Steffens concerns head and multiple landmark detection and recognition in video images and employs motion information:

9 Q I didn't ask for whether there was any
10 probabilities used at this point. I was just
11 asking, does Steffen employ a spatiotemporal
12 model?

13 ATTORNEY EKLEM: Objection to form.

14 A Well, I mean, it's important to
15 distinguish things. So it's -- it's -- you want
16 to be very clear. Steffen has essentially -- if
17 you look at Figure 1 of Steffen, you have a video
18 RAM 28 coming in, so you have video imagery coming
19 in; you have head detection and tracking in 14,
20 right, using a head detection and tracking
21 process; and then you have preselector; you have a
22 landmark finding that we've been talking about,

1 and you have a landmark tracking. And then you
2 have an identifier.
3 So essentially you're doing -- you know,
4 like if you look at column 3, I'm on line 57,
5 "Depending on the nature of the image data, the
6 head detection model uses one of a variety of
7 visual pathways which are based on, for example,
8 motion, color, or size (stereo vision) topology or
9 pattern."
10 So it is potentially employing some motion
11 information, but that's completely different from
12 what's being discussed in the '980. It's not --
13 it's not an apple-to-apple discussion or
14 comparison.

42. Further, Dr. Saber takes an inconsistent position that a POSITA would take as to the teachings of Steffens. For example, Dr. Saber acknowledges that Steffens discloses coarse “landmarks” and higher resolution “landmarks,” but draws an artificial distinction between “components” and “landmarks.” *See* Ex1051 at 253:8-266:10. Dr. Saber makes the same artificial distinction between similarity and probability, arguing that “a similarity measure [is] not a probability,” even though the more similar two vectors are, the higher their dot product and the closer to 1 the scalar product would be “if [] normalized.” *See* Ex1051 at 253:8-266:10. In my opinion, just because Steffens does not use identical terminology as

the '980 Patent does not mean a POSITA would not understand that the same concepts are disclosed in Steffens as that claimed by the purported invention of the '980 Patent. Here, a POSITA would understand Steffens' disclosure about "landmarks" and "similarity" to be discussing "components" and "probability" as used in the '980 Patent, as I detail below.

43. As stated in Steffens and in my First Declaration (*e.g.*, ¶¶55-59) and summarized here, Steffens indeed 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.

44. Also as stated in my First Declaration (¶¶135-136) and repeated here, Steffens discloses that its **model graph**, illustrated in Figure 12 with nodes corresponding to a collection of components, is spatio-temporal and designed for detecting and tracking a human face in video:

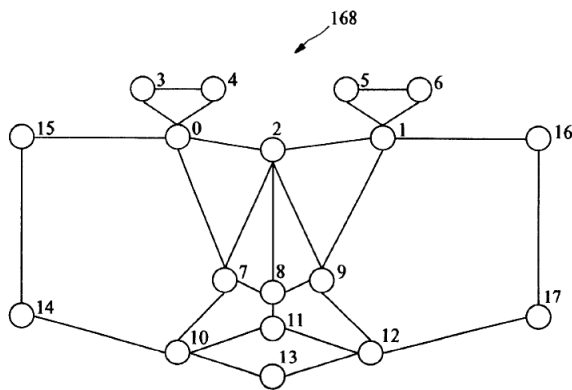
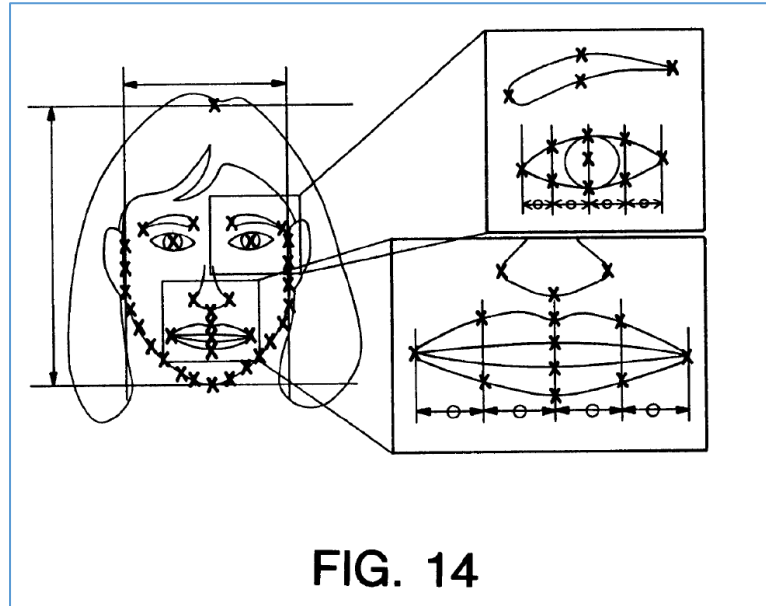


FIG. 12

- 0 right eye pupil
- 1 left eye pupil
- 2 top of the nose
- 3 right corner of the right eyebrow
- 4 left corner of the right eyebrow
- 5 right corner of the left eyebrow
- 6 left corner of the left eyebrow
- 7 right nostril
- 8 tip of the nose
- 9 left nostril
- 10 right corner of the mouth
- 11 center of the upper lip
- 12 left corner of the mouth
- 13 center of the lower lip
- 14 bottom of the right ear
- 15 top of the right ear
- 16 top of the left ear
- 17 bottom of the left ear

Ex1007 at 13 (Fig. 12).

45. To improve computational efficiency, Steffens (similar to Zhao) 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:

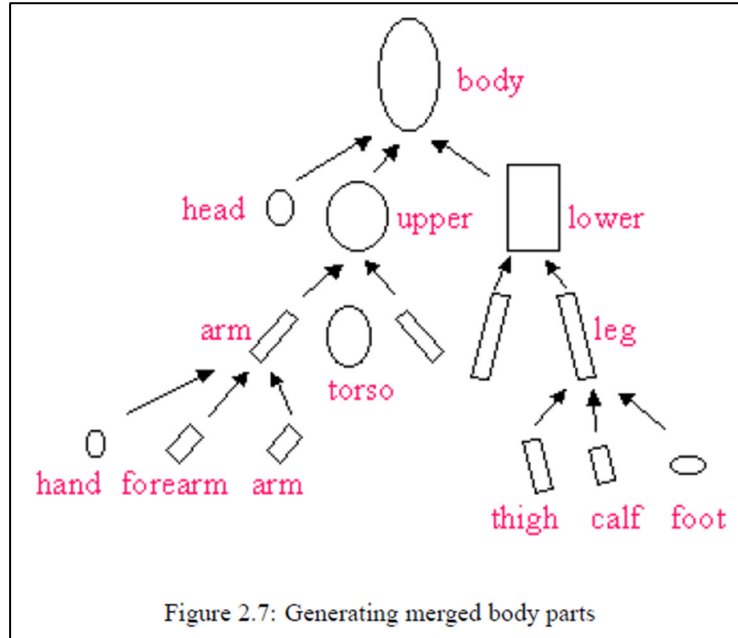


Ex1007 at 15 (Fig. 14)

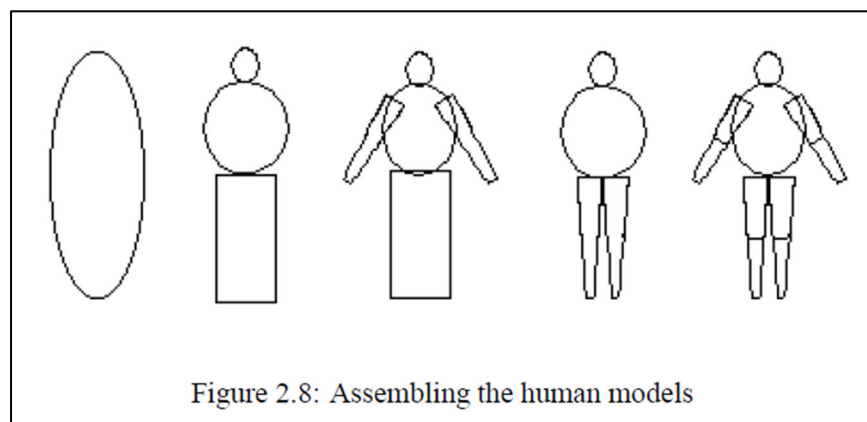
46. Regarding the “spatio” element, for its model for an object to be detected, Steffens says that “the nodes of the labeled graph refer to points on the object and are labeled by jets,” and “[e]dges of the graph are labeled with distance vectors between the nodes.” Ex1007 at 8:58-60. Further, “[n]odes 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.

47. Steffens further 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.

48. Notably, Zhao too 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”).

49. While Steffens focuses its object detection method on facial recognition using “confidence values” which “indicate the reliability of face detection and which are generated during bunch graph matching,” Steffens and Zhao each similarly concern object detection and recognition in video images based on multi-component, spatio-temporal graphical models and probabilistic matching. Zhao’s object model is a graphical model with nodes corresponding to an object and a collection of its components. Zhao explains (Ex1006 at 42), “[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.”

50. Similarly, Steffens defines a bunch graph, as illustrated in Figure 11. A bunch graph 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. Moreover, in Steffens, a model graph that is designed for finding a human face is illustrated in Figure 12 with nodes corresponding to a collection of components.

51. Dr. Saber states in ¶130 of his declaration that “Zhao teaches away from combining Steffen’s wavelet-based bunch graph approach with Zhao.” I disagree. Dr. Saber’s conclusion is apparently based on a misunderstanding of

normalized similarity measures and their relationship to probabilities. When asked during his deposition whether Figure 21 of Steffens illustrates matching and tracking of a face (Ex1051 at 262:15-264:18), Dr. Saber incorrectly concludes that Steffens' similarity measure "has nothing to do with probability at all."

15 Q Does Figure 21 of Steffen illustrate
16 matching and tracking of a face?

17 ATTORNEY EKLEM: Objection to form.

18 A So Figure 21, you see it has a
19 transformation into the Gabor space, which is
20 essentially like a wavelength decomposition, and
21 then you have a frontal bunch graph matching,
22 essentially, so you have graph matching each

1 time -- do you see how he's got his figure and you
2 see that little white stuff on the figure?

3 Q Yes.

4 A And you see how it goes to the next one,
5 and as it goes to the one on the right? So you
6 have the initial graph and the final graph, you
7 get that SF, the similarity has gone up to .74.
8 And when you go to the next one in the rotations,
9 you see the graph, and the similarity goes up to
10 .96. And then that's for the left one and for the
11 right one.

12 So he's essentially doing a similarity
13 measure that's based on the scale of product.
14 There is no discussion of probabilities here at
15 all.

16 And you look at the formula, it will tell
17 you exactly that. So if you look at the formula
18 for -- look on column 9, line 55 or 50 -- or 63,
19 whatever. Either one of them. So he said, "The
20 similarity of two jets" -- so I'm reading the
21 sentence above it: "If the components of the
22 Jet G are written in the form with amplitude AJ

1 and phase ϕ , the similarity of two jets, J and
2 J' , is the normalized scalar product of the
3 amplitude vector."
4 So you see $S(J, J')$ is equal to the scalar
5 product of $A(J, A, J')$ sum. And to normalize it
6 so it doesn't go beyond 1, he divides it by the
7 total sum of both of them, the total sum of the
8 square of the product of both of them. So this
9 way it's normalized.
10 But that's a similarity. That's like a
11 dot product. It's a scalar product, essentially.
12 That has nothing to do with probability at all.
13 Q What if the similarity is 1?
14 A Similarity has nothing do with
15 probability. There is no probability discussion
16 here. There is no probability in this whole
17 patent. I bet if you search on the word
18 "probability," it's not there.

52. As I explained in my First Declaration (§144), Steffens uses a similarity measure to match and determine whether an object is in an image. Steffens describes two different similarity functions that are both normalized so take on real similarity values ranging between $[0, 1]$ with 0 being not-similar and 1 being exactly similar, and real values between them giving a continuous measure function of similarity. As a POSITA would know this corresponds to a measure of probability of similarity between two entities. Ex1007 at 9:47-65.

If the graphs have relative distortion, a second term that accounts for geometrical distortions may be introduced. Two different jet similarity functions are used for two different, or even complementary, tasks. If the components of a jet $\vec{J}$ are written in the form with amplitude a_j and phase ϕ_j , the similarity of two jets $\vec{J}$ and $\vec{J}'$ is the normalized scalar product of the amplitude vector:

$$s(\vec{J}, \vec{J}') = \frac{\sum a_j a'_j}{\sqrt{\sum a_j^2 \sum a'_j^2}} \quad (4) \quad 55$$

The other similarity function has the form 60

$$s(\vec{J}, \vec{J}') = \frac{\sum a_j a'_j \cos(\phi_j - \phi'_j - \vec{d} \cdot \vec{k}_j)}{\sqrt{\sum a_j^2 \sum a'_j^2}} \quad (5)$$

53. With reference to Figure 21, Steffens discusses (Ex1007 at 14:41-61) a landmark finding process which attempts to match each bunch graph model (from a family of related bunch graphs) to an input image in order to determine the pose or size of the head in the image. According to Steffens, the graph that best corresponds to the pose given on the image has the highest similarity. Therefore a POSITA would understand that Steffens teaches a probabilistic measure of similarity of a bunch graph model that best corresponds to the pose given in the image through its normalized similarity functions.

54. Similar to Steffens, 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. Zhao’s contour-based approach is implemented using a Bayesian probabilistic framework and Steffen’s wavelet-based approach can be implemented using a probabilistic framework for object detection in lieu of a confidence-based similarity measure.

55. Applying a Bayesian (probabilistic) framework such as taught by Zhao to Steffens’s matching technique would have been obvious to a POSITA. A POSITA would have been motivated to use Zhao’s probabilistic framework to perform Steffen’s similarity analyses in graph matching since Zhao was also using normalized similarity measures ranging between $[0,1]$. *See* Ex1004, ¶125.

56. Moreover, Dr. Saber’s confusion about probabilities further extended to his interpretation of Zhao’s Bayesian probability similarity measure. When directly asked about the numbers stated above Figure 3.1. of Zhao, Dr. Saber first denied that Zhao taught detection probabilities of the object by itself in the image and also as a collection of components. Ex1006, Figure 3.1; Ex1051 at 232:13-235:12.

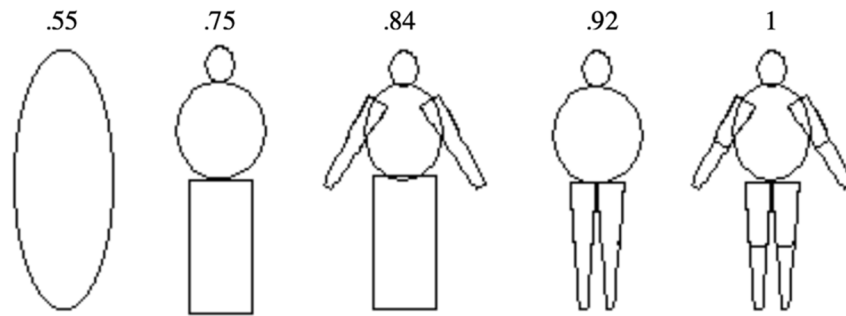


Figure 3.1: Examples of the calculated $P(\text{person}|H)$ given the degeneration factor $\alpha = .9$ and the contour parts matching the model parts exactly

13 So in Figure 3.1, each drawing, if you
14 will --
15 A Yes.
16 Q -- there are five drawings --
17 A Yes.
18 Q -- they all reflect a single person, which
19 is the object, correct?
20 A Well, it's the grouping of object parts
21 where the first one just has one thing and the
22 next one has three and the next one has five. So

1 just essentially it increases as you go all the
2 way to the right side, where you have the full
3 human body.
4 Q And the probability that the left drawing,
5 the oval, is a person is .55, correct?

17	So this .55, .75, whatever, is kind of
18	like a similarity measure. So in essence it
19	represents a probability, but it's in the spatial
20	domain. There is no temporal of any kind here.
21	This is all essentially spatial.
22	Q But it's the probability that the object

1	is in an image, correct?
2	A Well, it's -- essentially it's talking
3	about the contour and how well does the contour
4	matches that as, say, of a human, right?
5	Q Correct.

6 A So the contour that you're looking at, how
7 well does that match a person, if the person is
8 what I'm interested in? The closer it matches the
9 person, the higher this number gets. And the
10 lesser it matches a person -- like the ellipse on
11 the left, it obviously doesn't match a person very
12 well so he gave it a .55 weighting, indicating
13 that it's -- it's not very good, right? And 1
14 would be like it matches a person very well.

15 Q So doesn't it reflect that in the first
16 picture only one object is detected, the oval?

17 A Yes.

18 Q And therefore the probability that that
19 oval is a human is .55, correct?

20 A Generally speaking, yes.

1	Q Therefore there is a probability that the
2	object, including the three components, or
3	composed of the three components, the probability
4	that that is a human is .75, correct?
5	A Generally speaking, yes.
6	Q And then in the next picture, the system
7	has now detected five components, and therefore
8	the probability that that overall image is a human
9	is .84, and so on, correct?
10	A Yes. I mean, if you look at the caption
11	of the figure it says, "Examples of the calculated
12	probability of a person given the hypotheses."

57. To his credit and on further questioning, Dr. Saber additionally acknowledged that Zhao, like Steffens, resort to a similar coarse to fine procedure of matching based on probabilistic similarity measures, where the probability is computed for the object alone (first layer) and then for the collection of components (second layer). Hence as I stated in my First Declaration, Steffens' wavelet-based approach is not fundamentally different from or incompatible with Zhao's contour-based approach.

VI. A POSITA WOULD HAVE UNDERSTOOD THAT OZER'S "ACTIVITY" CLASSIFICATION ENCOMPASSES OBJECT TRACKING

58. Dr. Saber states in ¶150 of his declaration that:

Dr. Bajaj's reliance on Ozer's "Classifying Over Multiple Frames" is misplaced, as that part of the process is the separate "high level" process for classifying "activities," not the object.

And during his deposition (Ex1051 at 274:2-19), Dr. Saber mischaracterizes Ozer as just classifying activities (such as a human walking), and that not the same as tracking a moving object (the walking human) over multiple image frames.

2	Q	What does "activity" mean to you in the
3		context of the '266?
4	A	Let's say a human is walking, right?
5		You've got activity, right? The human is walking.
6		So you're looking at the different location of the
7		hands to say, okay, it's walking, it's an
8		activity.
9	Q	Over time?
10	A	But the detection of the object is done in
11		the frame. You look at the next frame, you can
12		see the motion of the hand from one point to the
13		other geometric features. There is no probability
14		here. There is no discussion of probabilities in
15		this case. There is no --
16	Q	Isn't the hidden Markov model a
17		probability?
18	A	Well, the hidden Markov model talks about
19		the different states of the activities, right?

59. As I stated in my First Declaration , my reliance on Ozer's "Classifying Activity Over Multiple Frames" is detecting and tracking the presence

of the moving object in multiple frames, is not misplaced. As I explained in ¶193 of my declaration, Ozer teaches (7:33-45):

The 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). {Emphasis added.}

60. In light of this passage from Ozer, among others, a POSITA would have understood that Ozer’s “activity” classification encompasses object motion tracking over multiple frames to classify or discriminate between various motion trajectories. Ozer teaches both the detection of the object parts and the movements (motions) of the object parts and moreover uses probabilistic Hidden Markov models for its inference of the motion (*e.g.*, walking activity) from video of the spatio-temporal model.

61. As I stated in my First Declaration (¶61), Ozer 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 that after the detection of the object parts, the system is ready to recognize the activities of each object part and the overall activity of the object. *Id.* at 4:35-37. Ozer “tracks only objects that fit the model of the user defined subject, such as human, any rigid object, dog.” *Id.* at 6:23-24. Ozer explains that “matched nodes” identified using the graph-matching algorithm are “tracked by using ellipse parameters for the consecutive frames.” *Id.* at 7:25-31.

64. A POSITA would have thus understood that Ozer classifies and tracks object parts over multiple frames.

VII. THE JOINT PROBABILITY DISTRIBUTION RECITED IN CLAIM 8/16 IS THE SAME PAIRWISE MARKOV FIELD REPRESENTATION OF A SPATIO-TEMPORAL GRAPHICAL MODEL DISCLOSED BY TLP AND PRIOR ART CITED IN MY FIRST DECLARATION

65. Dr. Saber provides multiple reasons (*e.g.*, ¶¶167-170) why he disagrees that the joint probability distribution formulated in claims 8/16 would have been well known to a POSITA and is also described by TLP “for its concept of non-parametric belief propagation.” Dr. Saber first argues (¶168) that “TLP does not disclose or render obvious a model with object-level nodes or object-level constraints, which are required by the probability distribution of claims 8/16.” And Dr. Saber further argues (¶169) that “TLP’s Figure 1 confirms the absence of object-level nodes and the absence of object level temporal constraints (directed edges).” I disagree with Dr. Saber’s arguments and elaborate why.

66. First, a joint probability distribution for the spatio-temporal model with N components is for any N, as shown in claims 8/16 (repeated below). Hence, and contrary to Dr. Saber’s statement, the lack of a certain specific node or having extra nodes does not affect the formula of the joint probability distribution.

$$P(X_0^O, X_0^{C_0}, X_0^{C_1}, \dots, X_0^{C_N}, \dots, X_T^O, X_T^{C_0}, X_T^{C_1}, \dots, X_T^{C_N}) = \frac{1}{Z} \prod_{ij} \psi_{ij}(X_i^O, X_j^O) \prod_{ik} \psi_{ik}(X_i^O, X_i^{C_k}) \prod_{ikl} \psi_{kl}(X_i^{C_k}, X_i^{C_l}) \prod_i \phi_i(X_i^O) \prod_{ik} \phi_i(X_i^{C_k}).$$

67. As stated in my First Declaration, this joint probability distribution is also disclosed in TLP (Ex1009 at 4) (“we use a form of non-parametric belief propagation [13 [Ex1019 PAMPAS], 25 [Ex1012 Sudderth]]”). I’ve included relevant excerpts from teachings of Ex1019, PAMPAS, that would have been well understood by a POSITA as teaching the same joint probabilistic distribution formulas of claim 8/16 and using the same belief propagation approach of the ’980 Patent.

Belief propagation can in general be stated using a pair-wise MRF formulation [26]. Consider a set of latent-variable nodes $\mathcal{X} = \{X_1, \dots, X_M\}$ paired with a set of observed nodes $\mathcal{Y} = \{Y_1, \dots, Y_M\}$. The conditional independence structure of the problem is expressed in the neighbourhood set $\mathcal{P}$. A pair of indices $(i, j) \in \mathcal{P}$ if the hidden node X_j is not conditionally independent of X_i given the other nodes in the graph, and in this case we say X_i is a parent of X_j and by slight abuse of notation $i \in \mathcal{P}(j)$.

For tutorial simplicity, when the X_i are discrete-valued random variables we assume each takes on some value $x_i \in \{1, \dots, L\} \equiv \mathcal{L}$. Given a fixed assignment to $\mathcal{Y}$ the observed and hidden nodes are related by observation

functions $\phi_i(x_i, y_i) \equiv \phi_i(x_i) : \mathcal{L} \rightarrow \mathbb{R}$. Adjacent hidden nodes are related by the correlation functions $\psi_{ij}(x_j, x_i) : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$. The joint probability over $\mathcal{X}$ is given by

$$P(\mathcal{X}) = \frac{1}{Z} \left(\prod_{(i,j) \in \mathcal{P}} \psi_{ij}(x_j, x_i) \right) \left(\prod_i \phi_i(x_i) \right). \quad (1)$$

The belief propagation algorithm introduces variables such as $m_{ij}(x_j)$ which can intuitively be understood as a *message* from hidden node i to hidden node j about what state node j should be in. Messages are computed iteratively using the update algorithm

$$m_{ij}(x_j) \leftarrow \sum_{x_i \in \mathcal{L}} \left(\phi_i(x_i) \psi_{ij}(x_j, x_i) \prod_{k \in \mathcal{P}(i) \setminus j} m_{ki}(x_i) \right) \quad (2)$$

and the marginal distribution of X_i (the node's *belief*) is given by

$$b_i(x_i) \propto \phi_i(x_i) \prod_{j \in \mathcal{P}(i)} m_{ji}(x_i). \quad (3)$$

68. I have also included relevant excerpts from teachings of Ex1012, Sudderth, that would have been well understood by a POSITA as teaching the same joint probabilistic distribution formulas of claim 8/16 and using the same belief propagation approach of the '980 Patent.

2. Undirected Graphical Models

An undirected graph $\mathcal{G}$ is defined by a set of nodes $\mathcal{V}$, and a corresponding set of edges $\mathcal{E}$. The *neighborhood* of a node $s \in \mathcal{V}$ is defined as $\Gamma(s) \triangleq \{t \mid (s, t) \in \mathcal{E}\}$. Graphical models associate each node $s \in \mathcal{V}$ with an unobserved, or hidden, random variable x_s , as well as a noisy local observation y_s . Let $x = \{x_s \mid s \in \mathcal{V}\}$ and $y = \{y_s \mid s \in \mathcal{V}\}$ denote the sets of all hidden and observed variables, respectively. For simplicity, we consider models with pairwise potential functions, for which $p(x, y)$ factorizes as

$$p(x, y) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{s,t}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s, y_s) \quad (1)$$

However, the nonparametric updates we present may be directly extended to models with higher-order potentials.

2.1. Belief Propagation

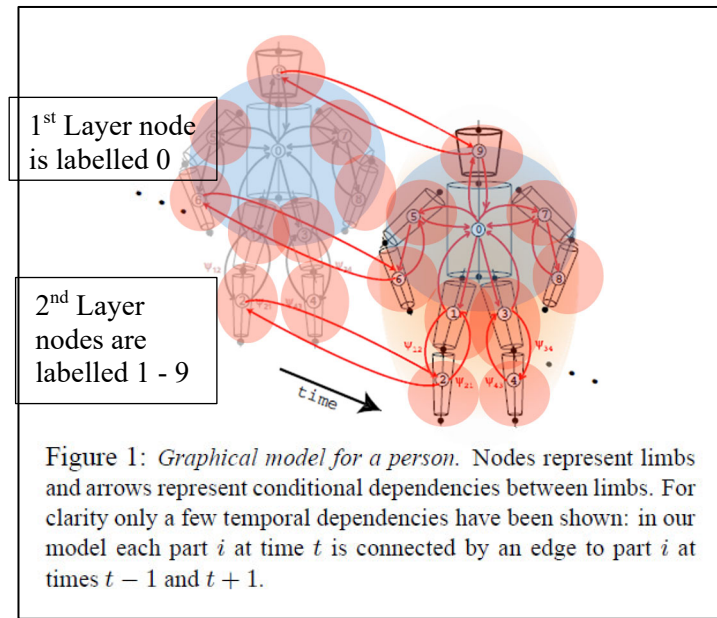
For graphs which are acyclic or tree-structured, the desired conditional distributions $p(x_s \mid y)$ can be directly calculated by a local message-passing algorithm known as *belief propagation* (BP) [16, 24]. At iteration n of the BP algorithm, each node $t \in \mathcal{V}$ calculates a message $m_{ts}^n(x_s)$ to be sent to each neighboring node $s \in \Gamma(t)$:

$$m_{ts}^n(x_s) = \alpha \int_{x_t} \psi_{s,t}(x_s, x_t) \psi_t(x_t, y_t) \times \prod_{u \in \Gamma(t) \setminus s} m_{ut}^{n-1}(x_t) dx_t \quad (2)$$

Here, α denotes an arbitrary proportionality constant. At any iteration, each node can produce an approximation $\hat{p}^n(x_s \mid y)$ to the marginal distribution $p(x_s \mid y)$ by combining the incoming messages with the local observation:

$$\hat{p}^n(x_s \mid y) = \alpha \psi_s(x_s, y_s) \prod_{t \in \Gamma(s)} m_{ts}^n(x_s) \quad (3)$$

69. 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 (Figure 1 Annotated).

70. Despite my discussion of TLP in my First Declaration, Dr. Saber testified during his deposition (Ex1051 at 111:2-21) that I provide no explanation how TLP determines the probability that an object is in each image:

1 So obviously I disagree. "Dr. Bajaj
2 provides no explanation as to how this,
3 'probabilistic method,' which determines a limb
4 likelihood, i.e., component likelihood, is
5 'equivalent,' to determining the probability that
6 the object is in each image."
7 The next paragraph is where the analysis
8 begins.
9 Q I'm sorry, please continue.
10 A So the next paragraph is where the
11 analysis begins. It says, "TLP makes clear to a
12 POSITA in its model, 'each node in the graphical
13 model represents...' " it's a whole description
14 here, "'a limb,'" we show that in Figure 1 in TLP,
15 "i.e., component," and this quote is from
16 Exhibit 9 at 1.
17 And the next paragraph says, "A POSITA
18 would have understood that TLP's model does not
19 include nodes that represent the object, only
20 components referred to as limbs or parts."
21 So I think that answers your question.

71. Tracking in TLP 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. TLP indicates that the body is treated as a graphical model, where each node in the graph corresponds to an independently parameterized body part. 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* to estimate the belief distribution over the parameters which accounts for the constraints and the observations, identical to the '980 Patent. *Id.*

72. Despite all the evidence I discussed in my First Declaration and amplified above with references to the prior art, in particular TLP, Dr. Saber testified (Ex1051 at 65:9-66:14) that the '980 Patent introduced likelihood functions and temporal compatibility functions:

11 Q Weren't these likelihood computations you
12 just discussed well-known in the relevant time
13 frame?

14 ATTORNEY EKLEM: Objection to form.

15 A No, this is not -- I mean, if we go back
16 to -- if you look at column 3, towards the bottom
17 of column 3, I'm looking at line 47 where it talks
18 about graphical object models, so this is part of
19 their contribution. So they propose this
20 spatiotemporal model. They have these
21 probabilities. You see "probability of," etc.
22 They have -- excuse me, they have likelihood

1 functions and they have temporal compatibility
2 functions that they introduce there. These are
3 introduced by them. These are not found in the
4 field. These are introduced in this patent.

5 They introduce the temporal connectivity
6 compatibility functions, they introduce local
7 evidence functions, which are the phis that you
8 see there, and I denote the local evidence. And
9 that's part of their formulation.

10 And then they describe how to compute
11 these compatibility functions and how to formulate
12 their likelihood functions and how do it the
13 classification as you continue reading column 4,
14 5, and 6, and maybe 7.

73. A POSITA would have understood from TLP how to resolve the graphical inference problem using Non-parametric Belief Propagation (NBP) or the PAMPAS algorithm that allows the parameters of the individual parts to be modeled using *continuous-valued* random variables. Ex1009 at 4. The body is represented by a graphical model in which each graph node corresponds to a body part (upper leg, torso, etc.). See Ex1009, Figure 1. Each 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. Edges in the graphical model correspond to **spatial and angular** and **temporal** relationships between adjacent body parts, as illustrated in Figure 1. As is standard for graphical models, a POSITA would have understood that the variables in a node are conditionally independent of those in non-neighboring nodes, given the values of the node's neighbors. *Id.* at 2.

74. A POSITA and certainly an expert in the field should have been familiar with the non-parametric belief propagation for graphical models generally and the PAMPAS paper in particular, given it is this "PAMPAS: Real valued graphical models in computer vision" paper that is cited in the '980 Patent provisional and patent specification. Surprisingly, as acknowledged during his deposition (Ex1051 at 78:4-21), Dr. Saber was not familiar with PAMPAS and did

not remember PAMPAS's disclosure of a graphical model "off the top of his head."

4	Q Okay. I want to return our attention to
5	the discussion of -- am I saying it correct,
6	PAMPAS, P-A-M-P-A-S?
7	A That's good enough for me.
8	Q Before your work on this IPR, were you
9	familiar with PAMPAS?
10	A I mean, I see a lot of stuff. I read a
11	lot of papers, so I'm sure I've come across these
12	type of things at one point or other. It's
13	published in a good conference. So computer
14	vision and pattern recognition is a good
15	conference.
16	Q Does PAMPAS disclose a graphical model
17	joint probability distribution formula?
18	A I don't remember off the top of my head,
19	sitting here today, what exactly is disclosed in
20	that paper. I couldn't tell you that sitting here
21	today. I would have to see the paper.

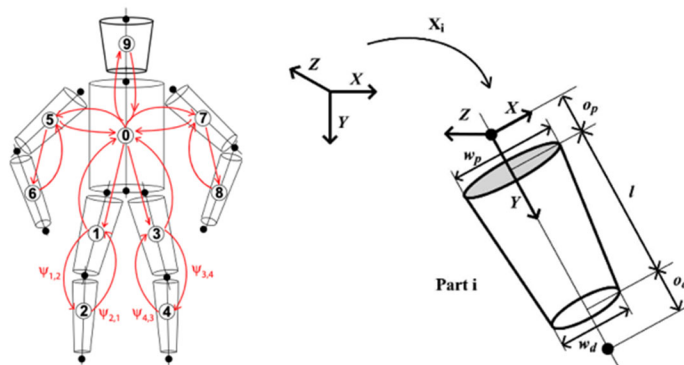
75. A POSITA would have understood from PAMPAS to select a component of the object (e.g., the torso) to be representative of the object node x_1 (i.e., a 1st layer), and to conditionally link all other components of the body (head, limbs, etc.) forming a 2nd layer of x_i nodes. See PAMPAS, Ex1019, § 3.3 Finding a jointed object in clutter. Due to this two-layered model, even when the object

node x_1 is occluded, after one iteration the belief at that node has high weight at the correct location due to the agreement of the nearby arm nodes x_i . After a few iterations the belief has converged on the correct object position despite the lack of the central circle. *Id.*

76. Moreover, the '980 Patent (6:13-22) itself claims scenarios where an object is modeled as a graphical collection of its components.

15 Depending on an object one may or may not have a likelihood or a proposal process for the object layer nodes. For example, if the whole appearance of an object is too complicated to model as a whole (e.g., arbitrary size vehicles) and can only be modeled in terms of components, a uniform likelihood can be assumed over the object state space. In such cases the object layer nodes simply fuse the component information to produce estimates for the object state that are consistent over time.

77. TLP (§2) teaches that each part/limb is modeled by a tapered cylinder having 5 fixed (person specific) and 6 estimated parameters.



Ex1009, Figure 1 Annotated; Figure 2 (Parameterization of part i). The fixed parameters $\Phi_i = (l_i, w_{ip}, w_{id}, o_{pi}, o_{di})$ correspond respectively to the part length,

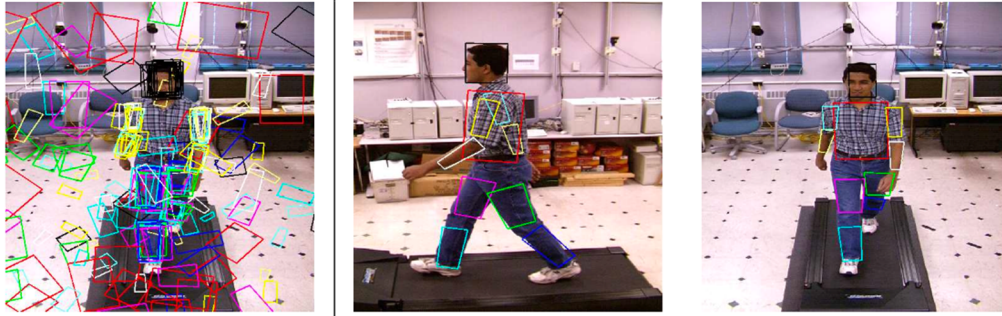
width at the proximal and distal ends and the offset of the proximal and distal joints along the axis of the limb as shown in Figure 2 of TLP. Each directed edge between parts i and j has an associated conditional distribution $\psi_{ij}(X_i, X_j)$ that encodes the compatibility between pairs of part configurations; that is, it models the probability of configuration X_j of part j conditioned on the X_i of part i .

78. Reference [22] cited by TLP, Attractive People: Assembling Loose-Limbed Models using Non-parametric Belief Propagation (“Attractive People”), Ex1015 (at 3), explains that:

For notational convenience we define an **ordering on body parts going from the torso out towards the extremities** and refer to **conditionals** that go along this ordering as **“forward” conditionals [directed edges]**. Conversely, the conditionals that go from the extremities towards the torso are referred to as “backward” conditionals [directed edges]. {Emphasis added.}

This terminology is identically used in the '980 Patent. *See* '980 Patent at 4:49-5:9.

79. According to Attractive People “[t]he inference algorithm ... combines the body model described above with a probabilistic image likelihood model.” Ex1015 at 4-5. An exemplary implementation of the inference algorithm is shown in Fig 4 for a human walking cycle with probabilistic likelihoods computed for the projected body parts that are visible in that video frame. *Id.* at 7. In particular, Attractive People teaches that $\phi_i(X_i)$ is the likelihood of observing the image measurements conditioned on the **pose of limb i** . *Id.* at 5.



Ex1015, Figure 4(a)

80. Thus, TLP does disclose or render obvious a joint probability distribution for a spatio-temporal model with N components of the same formula as claims 8/16.

81. Dr. Saber's second argument (§170) is that I have not shown that:

the joint probability formulation of claims 8/16 “would have been well known” or that Ex1012 discloses that formulation.... Dr. Bajaj identifies no disclosure in Ex1012 of object-level representation in combination with component-level representation.

I disagree.

82. As I explained in my declaration (§231), the equation recited in Claims 8/16 represent 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.

83. In support of my ¶231 opinion, my First Declaration cites Ex2012, a paper to Sudderth. Sudderth (at 2) explains:

The nonparametric belief propagation (NBP) algorithm developed in this paper extends the earlier work in two key ways. First, for graphs with cycles they do not form a junction tree, but instead iterate on local message updates until convergence as in loopy BP. This has the advantage of greatly reducing the dimensionality of the spaces over which they must infer distributions. In addition, they provide a message update algorithm specifically adapted to graphs containing continuous, non-Gaussian potentials. Each message is represented by a kernel-based nonparametric density estimate, and message products are found by an efficient local Gibbs sampling algorithm. They validate the NBP algorithm on a small Gaussian network, and present stereo vision results demonstrating its effectiveness. {Italics in original.}

84. A comparison of § 2.1 Belief Propagation of Sudderth and col. 6, lines 23-41 of the '980 reveals an identical disclosure including joint probability distributions. Moreover, as stated in ¶¶35 and 36 of my First Declaration and as repeated here:

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.”

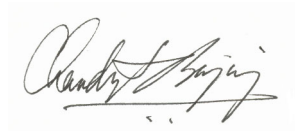
85. Thus, a POSITA would have understood the joint probability distribution of TLP to include object-level representation in combination with component-level representation.

VIII. CONCLUSION

For the foregoing reasons and the reasons I provided in my First Declaration, 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: June 17, 2024

Respectfully submitted,

A handwritten signature in black ink, appearing to read "Chandrajit Bajaj", is positioned above a horizontal line.

Chandrajit Bajaj, Ph.D.