

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
12 September 2008 (12.09.2008)

PCT

(10) International Publication Number  
**WO 2008/106725 A1**

(51) International Patent Classification:

*G06K 9/00* (2006.01)      *G06K 9/62* (2006.01)  
*G06K 9/74* (2006.01)      *G06T 15/00* (2006.01)  
*G06T 17/00* (2006.01)

(81) Designated States (unless otherwise indicated, for every

kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(21) International Application Number:

PCT/AU2008/000290

(22) International Filing Date: 4 March 2008 (04.03.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
2007901129      5 March 2007 (05.03.2007)      AU

(84) Designated States (unless otherwise indicated, for every

kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(71) Applicant (for all designated States except US): SEEING MACHINES PTY LTD [AU/AU]; Innovations Building, Corner Garren & Eggleston Road, Acton, ACT 2601 (AU).

(72) Inventor; and

(75) Inventor/Applicant (for US only): TELL, Dennis [SE/CH]; Ringstrasse 23, CH-8057 Zurich (CH).

Declaration under Rule 4.17:

— of inventorship (Rule 4.17(iv))

(74) Agent: SHELSTON IP; 60 Margaret Street, Sydney, NSW 2000 (AU).

Published:

— with international search report

(54) Title: EFFICIENT AND ACCURATE 3D OBJECT TRACKING

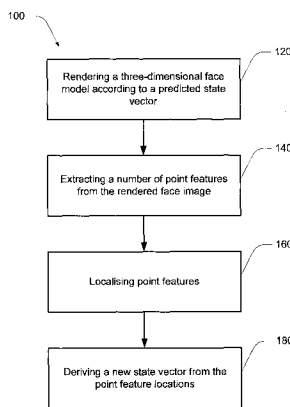


FIG. 1

(57) Abstract: A method of tracking an object in an input image stream, the method comprising iteratively applying the steps of: (a) rendering a three-dimensional object model according to a previously predicted state vector from a previous tracking loop or the state vector from an initialisation step; (b) extracting a series of point features from the rendered object; (c) localising corresponding point features in the input image stream; (d) deriving a new state vector from the point feature locations in the input image stream.



WO 2008/106725 A1

## EFFICIENT AND ACCURATE 3D OBJECT TRACKING

### FIELD OF THE INVENTION

[0001] The present invention relates to methods for measurement of the three-dimensional (3D) pose of a three-dimensional object.

- 5 [0002] The invention has been developed primarily for use in the context of human face tracking and will be described hereinafter with reference to this application. However, it will be appreciated that the invention is not limited to this particular field of use.

### BACKGROUND OF THE INVENTION

10 [0003] Any discussion of the prior art throughout the specification should in no way be considered as an admission that such prior art is widely known or forms part of common general knowledge in the field.

[0004] The tracking of faces and facial features, such as a person's eyes, has attracted considerable interest over the past decade, as computers have become sufficiently powerful to enable practical solutions to this problem.

- 15 [0005] There are two known approaches that have been proposed for solving this problem. The first approach is a geometric approach utilising three-dimensional point features in the face and geometric reasoning to derive the three-dimensional pose. The second is a non-linear optimisation of the parameters of an appearance model.

20 [0006] The first approach using point features has an advantage that it is a deterministic. Non-iterative approaches provide for short and predictable time to calculate the solution, and have been popular for real-time systems. *Edwards et al* (US patent no. 7,043,056) disclose a typical proponent of this methodology.

25 [0007] The second approach is an iterative, non-linear optimisation problem, which in general is computationally expensive. Trade-offs in terms of predictable convergence accuracy are required to achieve predictable computation times. Depending on the parameters of the appearance mode, an advantage of this approach is that a better fidelity of tracking can be achieved. It is understood that because the appearance of the observed object can be modelled and predicted more accurately than with the point

feature approach. *Cootes et al 2001* (T. Cootes, G. Edwards, C. Taylor, “Active appearance models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):681-685, 2001), *Cootes et al 2002* (T. Cootes, G. Wheeler, K. Walker, C. Taylor, “View-based active appearance models”, Image and Vision Computing, 20:657-664, 2002) and *Matthews et al* (I. Matthews and S. Baker, “Active appearance models revisited”, International Journal of Computer Vision, Vol. 60, No. 2, November, 2004, pp. 135 – 164) discloses a typical implementation of this methodology.

[0008] *De la Torre et al* (F. De la Torre, M. Black, “Robust parametrized component analysis: theory and applications of 2D facial appearance models”, Computer Vision and Image Understanding 91 (2003) 53-71) discloses the use of a person-specific two-dimensional active appearance model, which is not capable of tracking a person in three dimensions.

[0009] *Dornaika et al* (F. Dornaika, J. Ahlberg “Face model adaptation using robust matching and active appearance models”, Proceedings of Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002), 3-7) and *Hu et al* (C. Hu, R. Feris, M. Turk, “Active Wavelet Networks for Face Alignment”, Proceedings of British Machine Vision Conference, Norwich, 2003) disclose splitting a rendered face into sub-features, but fail to model and exploit the overlapping nature of facial features. A typical example for these occlusions occurs when the face is seen in a semi-profile view, where the ridge of the nose forms an edge over the far cheek. Previous systems are limited to the non-occluded view to features, and fail as soon as features start overlapping due to the projection of the facial features in the image.

[0010] It would be appreciated that technical challenges associated with this problem are considerable, in particular since an object such as the human face has high inter-individual variations, and the face is a highly articulate object.

[0011] There is a need in the art for automatic tracking of human faces in video sequences.

**SUMMARY OF THE INVENTION**

[0012] It is an object of the present invention to provide an improved form of three-dimensional object tracking.

[0013] In accordance with a first aspect of the present invention, there is provided a  
5 method of tracking an object in an input image stream, the method comprising  
iteratively applying the steps of: (a) rendering a three-dimensional object model  
according to a previously predicted state vector from a previous tracking loop or the  
state vector from an initialisation step; (b) extracting a series of point features from the  
rendered object; (c) localising corresponding point features in the input image stream;  
10 (d) deriving a new state vector from the point feature locations in the input image  
stream.

[0014] This method can also include an initialisation steps of (i) creation of a three-dimensional model of the object to be tracked; (ii) initial detection of the objects position within an input image stream.

15 [0015] Step (d) preferably includes the further step of: selecting a set of consistent features followed by Kalman filtering of the features to provide the new state vector.

[0016] The object can comprise a face and the state vector can comprise the three dimensional pose of the face. In one embodiment, the selecting of a set of features comprises random sample consensus (RANSAC) selection of the features. The  
20 localising step preferably can include utilising a normalised cross-correlation process to localise the point features.

[0017] It would be appreciated that embodiments of the present invention are has been primarily developed in the context of human face tracking. However, the present invention is applicable to other domains where the three-dimensional pose of an  
25 arbitrary object is to be measured visually.

[0018] In accordance with a second aspect of the present invention, there is provided a method of tracking an object in an input image stream, the method comprising steps of:

- (i) creating a three-dimensional model of the object to be tracked;

- 4 -

- (ii) localising initial features points in an initial input image stream;
- (iii) calculating an initial state vector indicative of the object location within the input image stream, wherein the initial state vector is calculated by minimising the square error between the initial localised feature points and corresponding initial feature points of the three-dimensional model projected into the image plane;
- 5
- (a) rendering a three-dimensional object model, wherein the object model accords with either the predicted state vector calculated in step (d) of a previous iteration or the initial state vector calculated in step (ii), wherein the rendering includes calculating a mask for the input image stream to distinguishing between background and foreground pixels;
- 10
- (b) calculating a predefined number of point features from the rendered object, wherein the predefined number of locations having highest edginess are selected as features from the rendered image of the previous iteration for the following localisation step;
- 15
- (c) localising corresponding point features in the input image stream;
- (d) calculating a new state vector from the localised point features in the input image stream; and
- (e) iteratively performing steps (a) though (d) for providing at each iteration updated the new state vector from the localised point features.
- 20

[0019] The localisation step preferably uses a masked normalised template correlation to calculate new the localised point features.

[0020] The new state vector is preferably calculated by using Random Sample Consensus selection and followed by a Kalman filter for estimating the new state vector. The Kalman filter preferably utilises a constant acceleration motion model of the object to predict the new state vector of the object from previously calculated state vectors and the new localised point features.

25

[0021] In accordance with a third aspect of the present invention, there is provided a computer program or computer program product comprising a set of instructions for carrying out a method according to any previously described method of tracking an object in an input image stream.

30

[0022] In accordance with a fourth aspect of the present invention, there is provided a computer-readable carrier medium carrying a set of instructions that when executed by one or more processors cause one or more processors to carry out a method of tracking an object in an input image stream, as previously described.

- 5 [0023] The carrier medium is preferably a medium bearing a propagated signal detectable by at least one processor of the one or more processors and representing the set of instructions.

[0024] In accordance with a fifth aspect of the present invention, there is provided a system for tracking an object in an input image stream, the system comprising a processor adapted to receive an input image stream, the processor is further adapted to perform a method of tracking an object in an input image stream, as previously described.

[0025] In accordance with a sixth aspect of the present invention, there is provided a system for tracking an object in an input image stream, the system comprising a processor adapted to receive an input image stream, the processor is further adapted to perform the steps of:

- (i) creating a three-dimensional model of the object to be tracked;
- (ii) localising initial features points in an initial input image stream;
- (iii) calculating an initial state vector indicative of the object location within the input image stream, wherein the initial state vector is calculated by minimising the square error between the initial localised feature points and corresponding initial feature points of the three-dimensional model projected into the image plane;
- (a) rendering a three-dimensional object model, wherein the object model accords with either the predicted state vector calculated in step (d) of a previous iteration or the initial state vector calculated in step (ii), wherein the rendering includes calculating a mask for the input image stream to distinguishing between background and foreground pixels;
- (b) calculating a predefined number of point features from the rendered object, wherein the predefined number of locations having highest edginess are

- 6 -

selected as features from the rendered image of the previous iteration for the following localisation step;

- (c) localising corresponding point features in the input image stream;
- (d) calculating a new state vector from the localised point features in the input image stream; and
- (e) iteratively performing steps (a) through (d) for providing at each iteration updated the new state vector from the localised point features.

[0026] The processor is preferably adapted to perform a masked normalised template correlation for calculating new localised point features. The processor is preferably adapted to apply the new localised point features into a Kalman filter for calculating the new state vector.

[0027] The invention has particular application in human performance assessment, the evaluation of the ergonomics of the design of human-machine interfaces, in safety applications to monitor distraction and fatigue, and in entertainment applications, where measurements are taken from a real person to be then rendered by a computer on an avatar.

#### **BRIEF DESCRIPTION OF THE DRAWINGS**

[0028] A preferred embodiment of the invention will now be described, by way of example only, with reference to the accompanying drawings in which:

- FIG. 1 is a flowchart of a method according to an embodiment of the invention;
- FIG. 2 is a flowchart of a method similar to Fig. 1, showing initialisations steps;
- FIG. 3 is a schematic view of a system according to an embodiment of the invention; and
- FIG. 4 is a schematic view of a system according to an embodiment of the invention, shown tracking a three-dimensional object.

#### **DESCRIPTION OF THE PREFERRED AND OTHER EMBODIMENTS**

[0029] The proposed method incorporates the advantages of both high accuracy in predicted appearance (and therefore high fidelity), and fast and predictable computation of point feature calculation. The preferred embodiment relates to methods for highly accurate yet efficient visual measurement of the three-dimensional

pose of a three-dimensional object. It would be appreciated that three-dimensional pose is a measure indicative of a three dimensional location and orientation, typically represented as six parameters. The method utilises components of feature based tracking methods with appearance-based methods referred to as “active appearance  
5 models”. The result is a highly efficient and accurate method to track and visually measure the pose of an object.

*Overview of the preferred embodiment*

[0030] Referring to FIG. 1, a flowchart representation 100 of method according to an embodiment of the present invention is disclosed, this method comprises the steps of:

- 10 (a) rendering a three-dimensional face model according to a predicted state vector; 120
- (b) extracting a number of point features from the rendered face image; 140
- (c) localising point features; 160 and
- (d) deriving a new state vector from the point feature locations 180.

15 [0031] Referring to FIG. 2, an initial step of a method 200, according to the preferred embodiment (and other related methods), includes creating a suitable three-dimensional model 210 of a face that is to be tracked - or equivalently an object that is to be tracked. Within this method, the makeup of the model is dependant on the type of tracking method. By way of example only, the model used in the method according  
20 to the preferred embodiment consists of a textured three-dimensional model of a face. This model allows a computer rendering of the face in any position and orientation. Locating a face within an image 215 provides an initial estimated facial location. Once the facial location is identified within an image, texture is mapped onto a generic three-dimensional model of a face by projecting the texture onto the vertices. By way of  
25 example only, *Loy et al* (Patent Cooperation Treaty, Application No. PCT/AU2003/000345, Publication Number WO/2003/081532) discloses a method used to identify the facial location within an image.

[0032] It would be appreciated by a person skilled in the art that the texture is mapped onto a generic three-dimensional model of a face by projecting the texture onto the  
30 vertices. By way of example only, a method of texture mapping is described in *Dornaika et al*.

[0033] After the above initialisation steps are performed, a main tracking loop is entered. This tracking loop utilises the textured three-dimensional model and estimates the state vector of the face for each image in an image stream. The steps involved in this tracking loop are described below:

- 5 (1) Rendering a three-dimensional face model according to a predicted state vector, 120. This predicted state vector is preferably derived from a previous tracking loop or a state vector from the initialisation steps.
- (2) Extracting a number of point features from the rendered face image, 140.
- (3) Localising point features, 160. Point feature localisation is derived with a  
10 suitable point feature localisation method. In the preferred embodiment the point feature localisation method is normalised cross-correlation.
- (4) Deriving a new state vector from the point feature locations 180. In the preferred embodiment, derivation of this new state vector from the point feature locations is performed through "Random Sample Consensus"  
15 (RANSAC) selection of a set of consistent features, and is followed by a Kalman filter to estimate the new state vector which in the present case is the three-dimensional pose.

[0034] This tracking loop, comprising steps (1) through (4) can be iterated indefinitely, or until the tracking of the face fails and a new initialisation is required.

20 [0035] By way of example only, individual steps are described in more detail below.

*Step (0) - Initialisation*

[0036] Initialisation requires the localisation of the face as a whole and key individual features. In the preferred embodiment, a method disclosed by *Loy et al* is utilised for the production of this localisation. This method allows the localisation of the face, the  
25 location of the four eye corners, the nostrils if visible and the two mouth corners.

These features are represented as reference points  $R = \{ r_i \in R^3 \}$  of the model.

[0037] Since this three-dimensional model uses a rigid three-dimensional model, only six pose parameters (for example including three translational and three rotational) are need to be estimated for fitting a generic three-dimensional face geometry model to the  
30 image. This generic model ' $M$ ' comprises corner vertices  $v_i$  and triangular planar

surfaces  $s_j$ , and approximates an average face geometry. This generic model can be expressed in the following equation.

$$M = \{ v_i \in R^3, s_j \}$$

[0038] It would be appreciated by a person skilled in the art that, representation of surface as a plurality of adjacent triangle is a technique used in three-dimensional graphic rendering and modelling. However, it would also be appreciated that other models are appropriate for use in the present invention.

[0039] By way of example only, this model captures a face from the forehead to the chin (typically in the vertical direction), and the area up to the ears (typically in the horizontal direction). This model excludes the hair of the person.

[0040] Three-dimensional pose  $P = \{ r_p \in R^3, t_p \in R^3 \}$ , comprising rotational components  $r_p \in R^3$  and translational components  $t_p \in R^3$ , is estimated by aligning the three-dimensional model with the localised image features. By way of example only, this alignment can be made such that the square-error between the reference points of the three-dimensional model projected into the image plane  $R(p)$  and the image locations of the image points  $R_i$  is minimised. This alignment minimisation can be represented by the following equation.

$$P_{init} = \underset{r,t}{\operatorname{argmin}} \left( \sum \| R(p) - R_i \|^2 \right)$$

[0041] *Dornaika et al* describes a method using a weak perspective assumption that is utilised in the preferred embodiment. It would be appreciated by a person skilled in the art that there are a number of optimisation techniques that can be used for performing model alignment.

[0042] Texture for each model triangle can be extracted from an image and stored with a three-dimensional geometric model. Combining the three-dimensional geometry and the texture for determining a three-dimensional model required for tracking. The

- 10 -

predicted head pose  $\bar{P}$  after initialisation is equivalent to the fitted three-dimensional model pose  $P_{init}$ .

*Step (1) - Rendering*

[0043] In the rendering step, the textured three-dimensional face model is typically rendered according to the head pose as predicted by the state estimation step (or the  
5 initialisation step if this is the first iteration). This can be computed using common computer graphics methods known to a person skilled in the art.

[0044] It would be appreciated that hardware (for example 3D graphics card) and/or software (for example OpenGL) can be used to accelerate this step. The person skilled in  
10 the art is aware of these techniques.

[0045] The output of this step includes a rendered face image and a mask for distinguishing between pixels that are background and pixels that are foreground.

*Step (2) - Feature extraction*

[0046] A set of features  $F = \{f_k\}$  are selected for use in the following steps. In the preferred embodiment, the set of features are dynamic and can vary at each tracking  
15 cycle to optimise robustness.

[0047] It would be appreciated that, by way of example only, feature extraction methods can use either an entire image rendered face image (for example as disclosed in *Cootes et al 2001* and *Cootes et al 2002*), vertices extracted from the model (for example as  
20 disclosed in *Dornaika et al*), or feature sets extracted from a frontal view without occlusions.

[0048] A drawback associated with using the entire rendered face is that the state vector can only be calculated in an iterative and computationally expensive manner. Depending on the distance metric used to establish the similarity of the rendered image  
25 with the live image from the camera, this approach typically has a significant sensitivity to inhomogeneous lighting conditions. Vertices or pre-extracted features can lack the ability to model and exploit the edges created from partial self-occlusion of the face due to the ridge of the nose occluding the far cheek when the head is rotated slightly.

[0049] To overcome these limitations, the method disclosed herein uses features computed from the rendered image in a previous tracking iteration. A predefined number of locations with the highest edginess are selected as features from the rendered image of the previous iteration for the following localisation step. The edginess 'e' of the location (x,y) in image 'I' is defined as the magnitude of the gradient of the image location, as calculated using the following equation.

$$e(I(x,y)) = \left| \frac{I(x,y)}{dxdy} \right|$$

[0050] An exhaustive search is performed to find  $f_k$  with the highest gradient values in the image. The output of this step includes a set of tracking features  $F = \{ f_k \}$ .

10 *Step (3) - Feature localisation*

[0051] Features selected in the previous step must be localised in a newly acquired image of the face. It would be appreciated by a person skilled in the art that this localisation could be computed using a number of methods used for feature localisation and tracking. In the preferred embodiment, a masked normalised template correlation is used to calculate the new location of the features.

[0052] For localisation, a 16x16 pixel template centred on the feature location  $f_k$  is extracted from the rendered image and masked by the rendering mask of the face, to exclude any background pixel which might be part of region defined by the 16x16 pixel template.

20 [0053] A person skilled in the art is aware of the commonly used formulation of normalised cross-correlation, as expressed in the following equation.

$$C(u,v) = \frac{\sum_{x,y} (I(u+x, v+y) - \bar{I}_{u,v}) \hat{T}(x,y)}{\sum_{x,y} (I(u+x, v+y) - \bar{I}_{u,v})^2 \sum_{x,y} \hat{T}(x,y)^2}$$

[0054] In this equation,  $I$  is the image and  $T$  is the template.  $\hat{T}$  is the brightness normalised template  $\hat{T} = T(x,y) - \bar{T}$  and  $\bar{T}$  is the template mean.  $\bar{I}_{u,v}$  is the local image mean at location (u,v), the mean of the image underneath the template whose

top left corner lies on pixel  $(u, v)$ . Furthermore, the sums and image mean are subject to the masking, i.e. not considering pixels that are masked out.

[0055] Around the predicted position of each feature, a search window is located.

Within this search window, an exhaustive search is performed to find the location of the feature with the aforementioned correlation technique. The location of that point  
5 (or pixel) with the highest correlation score is selected as the new feature location  $\hat{f}_k$ .

[0056] The output of the feature localisation step includes the new feature locations  $\hat{F}$  for the set of selected features  $F$ .

#### *Step (4) - State Estimation*

10 [0057] In the preferred embodiment, by way of example only, the state vector consists of the six three-dimensional pose-parameters only. This state vector can be calculated by applying the new feature locations into a Kalman filter of the type as disclosed by *Edwards et al.* This Kalman filter utilises a constant acceleration motion model of the face to predict the new position of the face from previous estimates and the newly  
15 acquired measurements. The Kalman filter can also allow the prediction of the head pose state vector  $\bar{P}$  for the next image frame.

[0058] The output of the state estimation step includes the predicted state vector  $\bar{P}$ , which is forwarded to the rendering step in the next iteration of the algorithm. It would be appreciated that steps (1) through (4) can be applied iteratively to provide a  
20 relatively efficient and accurate method for three-dimensional object tracking.

[0059] By way of example only, an embodiment of a method of tracking an object in an input image stream, comprises the steps of:

- (i) creating a three-dimensional model of the object to be tracked;
- (ii) localising initial feature points in an initial the input image stream;
- 25 (iii) calculating an initial state vector indicative of the object location within the input image stream, wherein the initial state vector is calculated by minimising the square error between the initial localised feature points and corresponding initial feature points of the three-dimensional model projected into the image plane;

- 5 (a) rendering a three-dimensional object model, wherein the object model accords with either the predicted state vector calculated in step (d) of a previous iteration or the initial state vector calculated in step (ii), wherein the rendering includes calculating a mask for the input image stream to distinguishing between background and foreground pixels;
- (b) calculating a predefined number of point features from the rendered object, wherein the predefined number of locations having highest edginess are selected as features from the rendered image of the previous iteration for the following localisation step;
- 10 (c) localising corresponding point features in the input image stream;
- (d) calculating a new state vector from the localised point features in the input image stream; and
- (e) iteratively performing steps (a) though (d) for providing at each iteration updated the new state vector from the localised point features.

15 [0060] It would be further appreciated that this method can be performed in real-time or on live video data.

[0061] It would be appreciated that the above methods can be performed, at least in part, by a computer program. A computer program or computer program product comprising a set of instructions for carrying out a method as previously described. A  
20 computer-readable carrier medium can embody the above methods. This computer-readable carrier medium carrying a set of instructions that when executed by one or more processors cause one or more processors to carry out a method according to a previously described method. This carrier medium is a medium bearing a propagated signal detectable by at least one processor of the one or more processors and  
25 representing the set of instructions.

[0062] FIG. 3 discloses a schematic view of a system according to an embodiment. A computer system 310, includes a processor 312 adapted to performer the steps of:

- (a) rendering a three-dimensional face model according to a predicted state vector;
- 30 (b) extracting a number of point features from the rendered face image;
- (c) localising point features; and

(d) deriving a new state vector from the point feature locations.

[0063] In an embodiment, the processor 312 is adapted to receive model data from a data storage device 314. This model data is indicative of a suitable three-dimensional model of a face (or object) that is to be tracked. It would be appreciated that, in an alternative embodiment, model data can be received from an external data storage device.

[0064] In an embodiment, the processor 312 is further adapted to receive live video data from a video capture device 320. It would be appreciated that, in an alternative embodiment, video data can be received from a data storage device (for example data storage device 314). The processor is adapted to locate a face (or object) within an image to provide an initial estimated facial location. Once the facial location is identified within an image, texture is mapped onto a generic three-dimensional model of a face by projecting the texture onto the vertices.

[0065] FIG. 4 discloses a schematic view of an embodiment shown tracking a three-dimensional object 410 in real time. In this embodiment a computer system 310 is couples to a video capture device in the form of a video camera 320. The video camera is adapted to transmit video data, which is received by the computer system. The computer system performs a method of measuring the three-dimensional (3D) pose of a three-dimensional object as previously disclosed. Data indicative of the measured three-dimensional pose can be output on an output devise 420. It would be appreciated that, in an alternative embodiment, data indicative of the measured three-dimensional pose can be stored in data storage device for later retrieval.

#### *Variations*

[0066] A number of variations of the above embodiment can include the use of more than one camera and the use of alternative methods employed in the individual computational steps.

[0067] Unless the context clearly requires otherwise, throughout the description and the claims, the words “comprise”, “comprising”, and the like are to be construed in an inclusive sense as opposed to an exclusive or exhaustive sense; that is to say, in the sense of “including, but not limited to”.

[0068] As used herein, unless otherwise specified the use of the ordinal adjectives “first”, “second”, “third”, etc., to describe a common object, merely indicate that different instances of like objects are being referred to, and are not intended to imply that the objects so described must be in a given sequence, either temporally, spatially,  
5 in ranking, or in any other manner.

[0069] Unless specifically stated otherwise, as apparent from the following discussions, it is appreciated that throughout the specification discussions utilizing terms such as “processing”, “computing”, “calculating”, “determining”, “applying”, “deriving” or the like, refer to the action and/or processes of a computer or computing system, or similar  
10 electronic computing device, that manipulate and/or transform data represented as physical, such as electronic, quantities into other data similarly represented as physical quantities.

[0070] In a similar manner, the term “processor” may refer to any device or portion of a device that processes electronic data, e.g., from registers and/or memory to transform  
15 that electronic data into other electronic data that, e.g., may be stored in registers and/or memory. A “computer” or a “computer system” or a “computing machine” or a “computing platform” may include one or more processors.

[0071] It will be understood that the steps of methods discussed are performed in one embodiment by an appropriate processor (or processors) of a processing (i.e.,  
20 computer) system executing instructions (computer-readable code) stored in storage. It will also be understood that the invention is not limited to any particular implementation or programming technique and that the invention may be implemented using any appropriate techniques for implementing the functionality described herein. The invention is not limited to any particular programming language or operating  
25 system.

[0072] It would be appreciated that, some of the embodiments are described herein as a method or combination of elements of a method that can be implemented by one or more processors of a computer system or by other means of carrying out the function. Thus, a processor with the necessary instructions for carrying out such a method or  
30 element of a method forms a means for carrying out the method or element of a method. Furthermore, an element described herein of an apparatus embodiment is an

- 16 -

example of a means for carrying out the function performed by the element for the purpose of carrying out the invention.

[0073] In alternative embodiments, the computer system comprising one or more processors operates as a standalone device or may be configured, e.g., networked to other processor(s), in a networked deployment. The one or more processors may operate in the capacity of a server or a client machine in server-client network environment, or as a peer machine in a peer-to-peer or distributed network environment.

[0074] Thus, one embodiment of each of the methods described herein is in the form of a computer-readable carrier medium carrying a set of instructions, e.g., a computer program that are for execution on one or more processors.

[0075] Reference throughout this specification to “one embodiment” or “an embodiment” means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment. Thus, appearances of the phrases “in one embodiment” or “in an embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment, but may refer to the same embodiment. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to one of ordinary skill in the art from this disclosure, in one or more embodiments.

[0076] Similarly it should be appreciated that in the above description of exemplary embodiments of the invention, various features of the invention are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure and aiding in the understanding of one or more of the various inventive aspects. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed invention requires more features than are expressly recited in each claim. Rather, as the following claims reflect, inventive aspects lie in less than all features of a single foregoing disclosed embodiment. Thus, the claims following the Detailed Description are hereby expressly incorporated into this Detailed Description, with each claim standing on its own as a separate embodiment of this invention.

[0077] Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention, and form different embodiments, as would be understood by those in the art. For example, in the  
5 following claims, any of the claimed embodiments can be used in any combination.

[0078] In the description provided herein, numerous specific details are set forth. However, it is understood that embodiments of the invention may be practiced without these specific details. In other instances, well-known methods, structures and techniques have not been shown in detail in order not to obscure an understanding of  
10 this description.

[0079] Although the invention has been described with reference to specific examples it will be appreciated by those skilled in the art that the invention may be embodied in many other forms.

## THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:-

1. A method of tracking an object in an input image stream, said method comprising iteratively applying the steps of:
  - (a) rendering a three-dimensional object model according to a previously predicted state vector from a previous tracking loop or a state vector from an initialisation step;
  - (b) extracting a series of point features from said rendered object;
  - (c) localising corresponding point features in said input image stream;
  - (d) deriving a new state vector from said localised point feature in the input image stream.
2. A method as claimed in claim 1 further comprising an initialisation step of
  - (i) creation of a three-dimensional model of said object to be tracked;
  - (ii) initial detection of said objects position within said input image stream.
3. A method as claimed in claim 1 wherein said step (d) further includes the step of selecting a set of consistent features followed by Kalman filtering of said features for deriving said new state vector.
4. A method as claimed in claim 1 wherein said object comprises a face.
5. A method as claimed in claim 4 wherein said state vector comprises the three-dimensional pose of said face.
6. A method as claimed in claim 3 wherein the selecting of a set of features comprises random sample consensus (RANSAC) selection of said features.
7. A method as claimed in claim 1 wherein said localising step includes utilising a normalised cross-correlation process to localise said point features.
8. A method of tracking an object in an input image stream, the method comprising steps of:
  - (i) creating a three-dimensional model of said object to be tracked;
  - (ii) localising initial features points in an initial said input image stream;

- (iii) calculating an initial state vector indicative of said object location within said input image stream, wherein said initial state vector is calculated by minimising the square error between said initial localised feature points and corresponding initial feature points of said three-dimensional model projected into the image plane;
- 5 (a) rendering a three-dimensional object model, wherein said object model accords with either said predicted state vector calculated in step (d) of a previous iteration or said initial state vector calculated in step (ii), wherein said rendering includes calculating a mask for said input image stream to distinguishing between background and foreground pixels;
- 10 (b) calculating a predefined number of point features from said rendered object, wherein said predefined number of locations having highest edginess are selected as features from said rendered image of the previous iteration for the following localisation step;
- (c) localising corresponding point features in said input image stream;
- 15 (d) calculating a new state vector from said localised point features in said input image stream; and
- (e) iteratively performing steps (a) though (d) for providing at each iteration updated said new state vector from said localised point features.
- 20 9. A method according to claim 8, wherein said localisation step uses a masked normalised template correlation to calculate new said localised point features.
10. A method according to claim 8 or claim 9, wherein said new state vector is calculated by using Random Sample Consensus selection and followed by a Kalman filter for estimating said new state vector.
- 25 11. A method according to claim 10, wherein said Kalman filter utilises a constant acceleration motion model of said object to predict said new state vector of said object from previously calculated state vectors and said new localised point features.
12. A method of tracking an object substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.
- 30

13. A computer program or computer program product comprising a set of instructions for carrying out a method according to any one of claims 1 to 12.
14. A computer-readable carrier medium carrying a set of instructions that when executed by one or more processors cause one or more processors to carry out a method according to any one of claims 1 to 12.
15. A carrier medium as recited in claim 14, wherein the carrier medium is a medium bearing a propagated signal detectable by at least one processor of said one or more processors and representing said set of instructions.
16. A system for tracking an object in an input image stream, said system comprising a processor adapted to receive an input image stream, said processor is further adapted to perform a method according to any one of claims 1 to 12.
17. A system for tracking an object in an input image stream, the system comprising a processor adapted to receive an input image stream, said processor is further adapted to perform the steps of:
- (i) creating a three-dimensional model of said object to be tracked;
  - (ii) localising initial features points in an initial said input image stream;
  - (iii) calculating an initial state vector indicative of said object location within said input image stream, wherein said initial state vector is calculated by minimising the square error between said initial localised feature points and corresponding initial feature points of said three-dimensional model projected into the image plane;
- (a) rendering a three-dimensional object model, wherein said object model accords with either said predicted state vector calculated in step (d) of a previous iteration or said initial state vector calculated in step (ii), wherein said rendering includes calculating a mask for said input image stream to distinguishing between background and foreground pixels;
  - (b) calculating a predefined number of point features from said rendered object, wherein said predefined number of locations having highest edginess are selected as features from said rendered image of the previous iteration for the following localisation step;
  - (c) localising corresponding point features in said input image stream;

- 21 -

- (d) calculating a new state vector from said localised point features in said input image stream; and
- (e) iteratively performing steps (a) through (d) for providing at each iteration updated said new state vector from said localised point features.

- 5 18. A system according to claim 17, wherein said processor is adapted to perform a masked normalised template correlation for calculating new said localised point features.
19. A system according to claim 17, wherein said processor is adapted to apply said new localised point features into a Kalman filter for calculating said new state
- 10 vector.
20. A system for tracking an object in an input image stream substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.

1/3

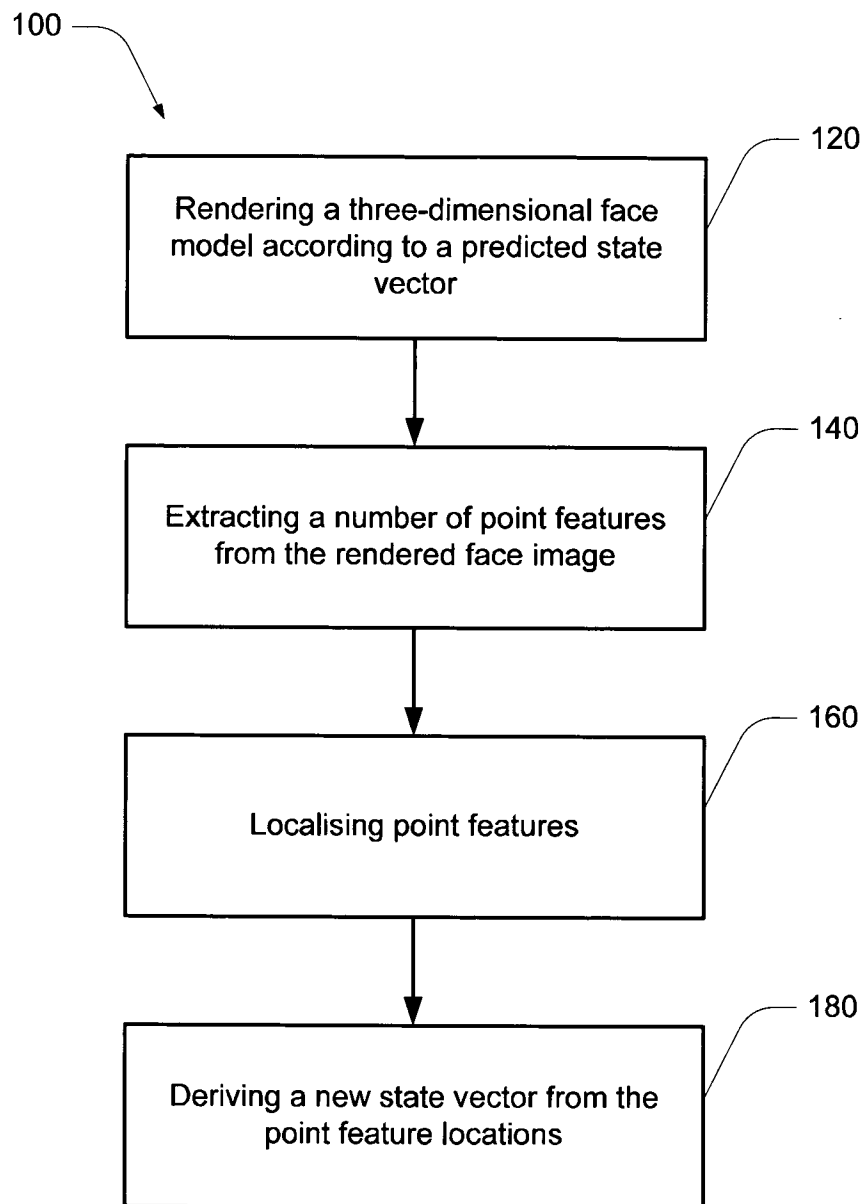


FIG. 1

2/3

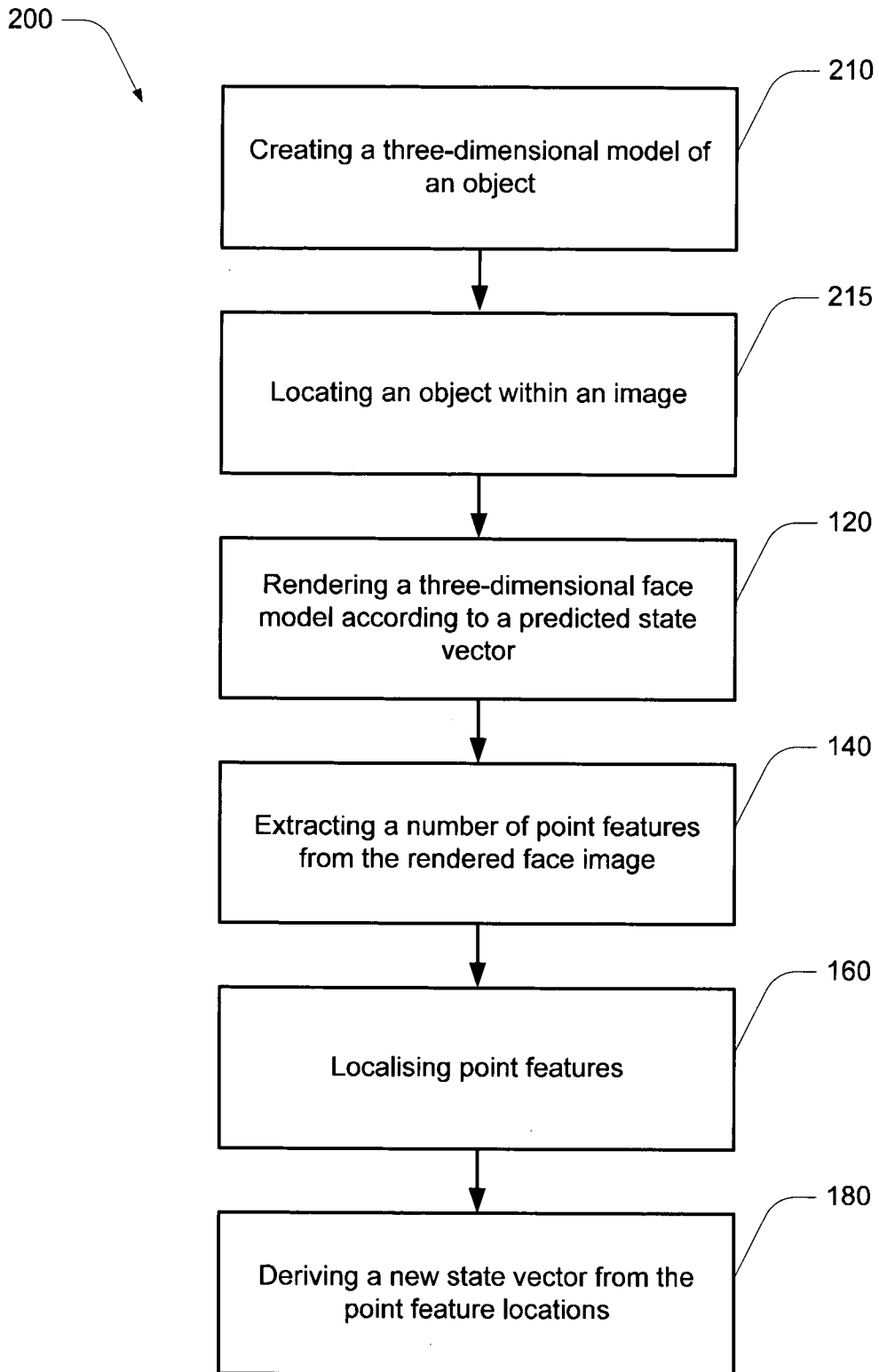


FIG. 2

3/3

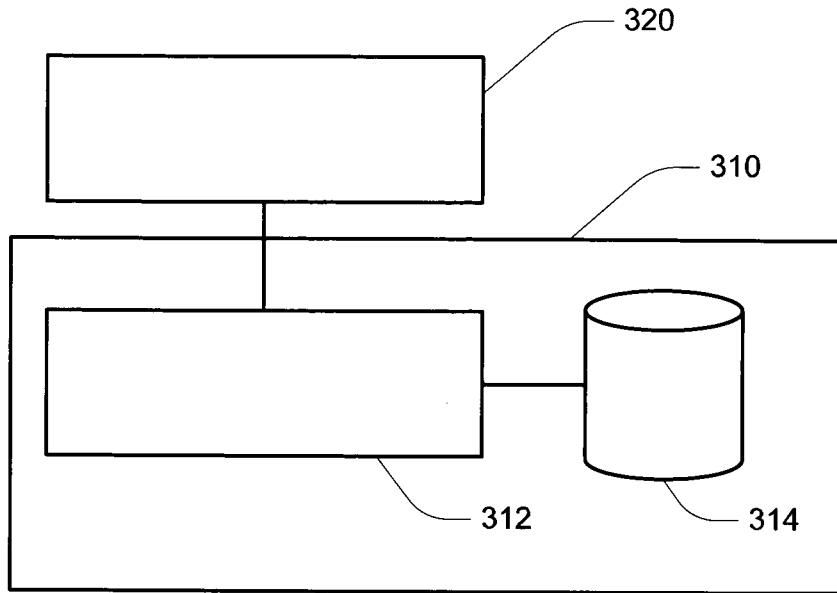


FIG. 3

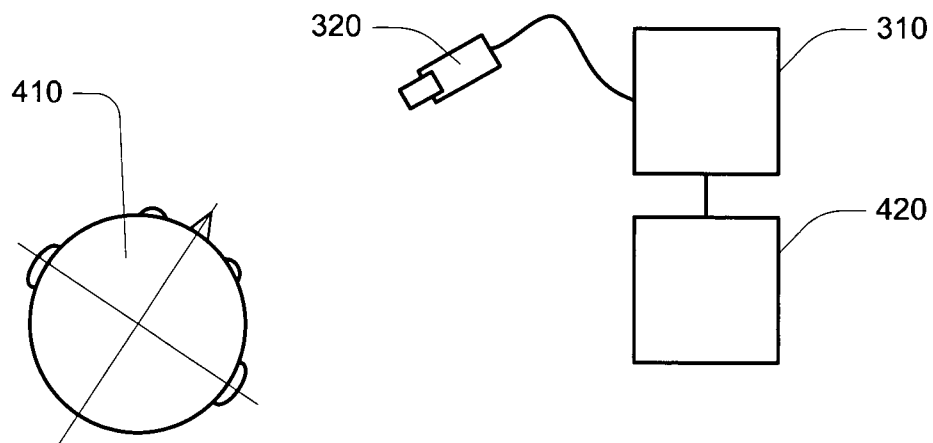


FIG. 4

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/AU2008/000290

## A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl.

G06K 9/00 (2006.01)

G06K 9/74 (2006.01)

G06T 17/00 (2006.01)

G06K 9/62 (2006.01)

G06T 15/00 (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Google Scholar, ESPACE, DWPI + G06k, G06T, G06F + Keywords: recogn+, track+, pose, represent+, monitor+, inspect+, video, render+, raster+, model+, wireframe?, point feature, vector?, state vector?, predict, +iterat+

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 20040165781 A (SUN) 26 August 2004 Entire document. See in particular: the abstract, drawings, paras [2], [7], [9] to [16], [37] to [45], [68] to [76], for example	1, 2, 4, 5
Y	Entire document	3, 6-20
Y	US 6917692 B (MURCHING et al.) 12 July 2005 Entire document.	3, 10, 11, 18
Y	EP 1139269 A (NEC CORPORATION) 4 October 2001 Entire document. See in particular: the abstract and drawings and pages 2-7 for example	6, 7, 10, 11, 18

 Further documents are listed in the continuation of Box C See patent family annex

* Special categories of cited documents:	
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E" earlier application or patent but published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

07 April 2008

Date of mailing of the international search report

16 APR 2008

Name and mailing address of the ISA/AU

AUSTRALIAN PATENT OFFICE  
PO BOX 200, WODEN ACT 2606, AUSTRALIA  
E-mail address: pct@ipaustralia.gov.au  
Facsimile No. +61 2 6283 7999

Authorized officer

**CHARLES BERKO**  
AUSTRALIAN PATENT OFFICE  
(ISO 9001 Quality Certified Service)  
Telephone No : (02) 6283

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/AU2008/000290

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2004/111927 A (UNIVERSITÉ LAVAL) 23 December 2004 Entire document. See in particular: pages 6-9 line 5 for example, and the drawings	1-20
Y	US 6256418 B (REHG et al.) 3 July 2001 Entire document. See in particular: the abstract, drawings, col. 1 line 39 to col. 2; col. 5 line 9 to col. 10 for example	1-20
Y	US 20040247174 A (LYONS et al.) 9 December 2004 Entire document. See in particular: the abstract, drawings, paras [49], [54], [161] to [187], [338], [365] to [439], [473] to [516] and claims for example	1-20
Y	Seeing Machines, Makers of faceLAB – Head, Eye & Blink tracking. Creators of Visionary Technology. faceAPI. Real time 3D head Tracker Demonstration. The Ultimate Real-Time face Tracking Toolkit for Developers and OEM's. Retrieved on 2 April 2008. <a href="http://www.archive.org/web/20070127214200/http://www.seeingmachines.com/">www.archive.org/web/20070127214200/http://www.seeingmachines.com/</a> ; <a href="http://www.seeingmachines.com/faceAPI.html">www.seeingmachines.com/faceAPI.html</a> ;	1-20
Y	Face Model adaptation using Robust matching and active appearance models. Dornaika and Ahlberg. 2002 (WACV) Proceedings Sixth IEEE Workshop on Applications of Computer Vision	1-20
Y	Fast and Reliable Active Appearance Model Search for 3-D Face Tracking. Dornaika and Ahlberg. IEEE Transactions on Systems, Man, and Cybernetics – part B CYBERNETICS, Vol 34. No. 4 August 2004	1-20
Y	The image Chip for High Performance 3D Rendering by Dunnett et al. Computer Graphics and Applications. IEEE Volume 12. Issue 6. Nov 1992 pp 41-52	1-20
A	Advances in Active Appearance Models. Cootes et al. The proceedings of the seventh IEEE International Conference on Computer vision. Volume 1. 1999 pages 137-142	1-20

**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/AU2008/000290**

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member			
US	2004165781	US	7164800		
US	6917692	NONE			
EP	1139269	JP	2001283222	US	6956569
WO	2004111927	CA	2529044	US	2007052974
US	6256418	NONE			
US	2004247174	GB	2358540	GB	2359686
		US	6980690	GB	2365243

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX