(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 May 2009 (14.05.2009)

PCT

(10) International Publication Number
WO 2009/061283 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/SG2008/000428

(22) International Filing Date:
7 November 2008 (07.11.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/002,627 9 November 2007 (09.11.2007) US

(71) Applicant *(for all designated States except US)*: **NATIONAL UNIVERSITY OF SINGAPORE** [SG/SG]; 21 Lower Kent Ridge Road, Singapore 119077 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LEOW, Wee Kheng** [SG/SG]; National University of Singapore,School of Computer Science, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG). **WANG, Ruixuan** [CN/SG]; National University of Singapore,School of Computer Science, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG). **LEE, Chee-Seng, Mark** [MY/US]; 8008 Glitter Court, Orlando, Florida 32836 (US). **XING, Dongfeng** [CN/SG]; National University of Singapore,School of Computer Science, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG). **LEONG, Hon Wai** [MY/SG]; National University of Singapore,School of Computer Science, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG).

(74) Agent: **ELLA CHEONG SPRUSON & FERGUSON (SINGAPORE) PTE LTD**; P.O. Box 1531, Robinson Road Post Office, Singapore 903031 (SG).

(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, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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),

[Continued on next page]

(54) Title: HUMAN MOTION ANALYSIS SYSTEM AND METHOD

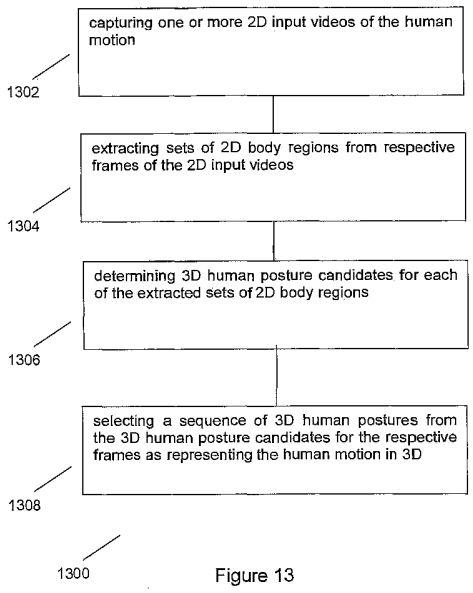

Figure 13

(57) Abstract: A method and system for human motion analysis. The method comprises the steps of capturing one or more 2D input videos of the human motion; extracting sets of 2D body regions from respective frames of the 2D input videos; determining 3D human posture candidates for each of the extracted sets of 2D body regions; and selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

WO 2009/061283 A2

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).

Published:

— *without international search report and to be republished upon receipt of that report*

Human Motion Analysis System and Method

FIELD OF INVENTION

The present invention relates broadly to a method and system for human motion analysis.

BACKGROUND

There are two general types of systems that can be used for motion analysis: 2D video-based software and 3D motion capture systems. 2D video-based software such as V1 Pro [V1 Pro, swing analysis software, www.v1golf.com] , MotionView [MotionView, golf swing video and motion analysis software, www.golfcoachsystems.com/golf-swing-software.htm], MotionCoach [MotionCoach, golf swing analysis system, www.motioncoach.com], and cSwing 2008 [cSwing 2008, video swing analysis program, www.cswing.com] provide a set of tools for the user to manually assess his performance. It is affordable but lacks the intelligence to perform the assessment automatically. The assessment accuracy depends on the user's competence in using the software. Such systems perform assessment only in 2D, which is less accurate than 3D assessment. For example, accuracy may be reduced due to depth ambiguity in 3D motion and self-occlusions of body parts.

3D motion capture systems such as Vicon [Vicon 3D motion capture system, www.vicon.com/applications/sports.html] and MAC Eagle [Motion Analysis Corporation, Eagle motion capture system, www.motionanalysis.com] capture 3D human motion by tracking reflective markers attached to the human body and computing the markers' positions in 3D. Using specialized cameras, these systems can capture 3D motion efficiently and accurately. Given the captured 3D motion, it is relatively easy for an add-on algorithm to compute the motion discrepancies of the user's motion relative to domain-specific reference motion. However, they are not equipped with an intelligent software for automatic assessment of the motion discrepancies based on domain-

specific assessment criteria. They are very expensive systems requiring six or more cameras to function effectively. They are also cumbersome to set up and difficult to use. These are passive marker-based systems.

There is also available an active marker-based system. In the system, the markers are LEDs that each blink a special code that uniquely identifies the marker. Such systems can resolve some tracking difficulties of passive marker-based system. However, the LEDs are connected by cables which supply electricity for them to operate. Such a tethered system places restriction on the kind of motion that can be captured. So, it is less versatile than untethered systems.

U.S. Patents US 4891748, US 7095388, disclose systems that capture the video of a person performing a physical skill, project the reference video of an expert scaled according to the body size of the person, and compare the motion in the videos of the person and the expert. In these systems, motion comparison is performed only in 2D videos. They are not accurate enough and may fail due to depth ambiguity in 3D motion and self-occlusions of body parts.

Japanese Patent JP 2794018 discloses a golf swing analysis system that attaches a large number of markers onto a golfer's body and club, and captures a sequence of golf swing images using a camera. The system then computes the makers' coordinates in 2D, and compares the coordinate data with those in a selected reference data.

US Patents US 2004/0209698 and US 7097459 disclose similar systems as JP 2794018 except that two or more cameras are used to capture multiple simultaneous image sequences. Therefore, they have the potential to compute 3D coordinates. These are essentially marker-based motion capture systems.

US Patent Publication US 2006/0211522 discloses a system of colored markers placed on a baseball player's arms, legs, bat, pitching mat, etc. for manually facilitating the proper form of the player's body. No computerized analysis and comparison is described in the patent.

US Patent US 5907819 discloses a golf swing analysis system that attaches motion sensors on the golfer's body. The sensors record the player's motion and send the data to a computer through connecting cables to analyze the player's motion.

Japanese Patents JP 9-154996, JP 2001-614, and European Patent EP 1688746 describe similar systems that attach sensors to the human body. US Patent Publication 2002/0115046 and US Patent 6567536 disclose similar systems except that a video camera is also used to capture video information which is synchronized with the sensor data. Since the sensors are connected to the computer by cables, the motion type that can be captured is restricted. These are tethered systems, as opposed to the marker-based systems described above, which are untethered.

US Patent US 7128675 discloses a method of analyzing a golf swing by attaching two lasers to the putter. A camera connected to a computer records the laser traces and provides feedback to the golfer regarding his putting swing. For the same reason as the methods that use motion sensors, the motion type that can be captured is restricted.

A need therefore exists to provide a human motion analysis system and method that seek to address at least one of the above-mentioned problems.

SUMMARY

In accordance with a first aspect of the present invention there is provided a method for human motion analysis, the method comprising the steps of capturing one or more 2D input videos of the human motion; extracting sets of 2D body regions from respective frames of the 2D input videos; determining 3D human posture candidates for each of the extracted sets of 2D body regions; and selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

The method may further comprise the step of determining differences between 3D reference data for said human motion and the selected sequence of 3D human postures.

The method may further comprise the step of visualizing said differences to a user.

Extracting the sets of 2D body regions may comprise one or more of a group consisting of background subtraction, iterative graph-cut segmentation and skin detection.

Determining the 3D human posture candidates may comprise the steps of generating a first 3D human posture candidate; and flipping a depth orientation of body parts represented in the first 3D human posture candidate around respective joints to generate further 3D human posture candidates from the first 3D human posture candidate.

Generating the first 3D human posture candidate may comprise temporally aligning the extracted sets of 2D body portions from each frame with 3D reference data of the human motion and adjusting the 3D reference data to match the 2D body portions.

Selecting the sequence of 3D human postures from the 3D human posture candidates may be based on a least cost path among the 3D human posture candidates for the respective frames.

Selecting the sequence of 3D human postures from the 3D human posture candidates may further comprise refining a temporal alignment of the extracted sets of 2D body portions from each frame with 3D reference data of the human motion.

In accordance with a second aspect of the present invention there is provided a system for human motion analysis, the method comprising the steps of means for capturing one or more 2D input videos of the human motion; means for extracting sets of 2D body regions from respective frames of the 2D input videos; means for determining 3D human posture candidates for each of the extracted sets of 2D body regions; and means for selecting a sequence of 3D human postures from the 3D

human posture candidates for the respective frames as representing the human motion in 3D.

The system may further comprise means for determining differences between 3D reference data for said human motion and the selected sequence of 3D human postures.

The system may further comprise means for visualizing said differences to a user.

The means for extracting the sets of 2D body regions may perform one or more of a group consisting of background subtraction, iterative graph-cut segmentation and skin detection.

The means for determining the 3D human posture candidates may generate a first 3D human posture candidate; and flips a depth orientation of body parts represented in the first 3D human posture candidate around respective joints to generate further 3D human posture candidates from the first 3D human posture candidate.

Generating the first 3D human posture candidate may comprise temporally aligning the extracted sets of 2D body portions from each frame with 3D reference data of the human motion and adjusting the 3D reference data to match the 2D body portions.

The means for selecting the sequence of 3D human postures from the 3D human posture candidates may determine a least cost path among the 3D human posture candidates for the respective frames.

The means for selecting the sequence of 3D human postures from the 3D human posture candidates may further comprise means for refining a temporal alignment of the extracted sets of 2D body portions from each frame with 3D reference data of the human motion.

In accordance with a third aspect of the present invention there is provided a data storage medium having computer code means for instructing a computing device to execute a method for human motion detection, the method comprising the steps of capturing one or more 2D input videos of the human motion; extracting sets of 2D body regions from respective frames of the 2D input videos; determining 3D human posture candidates for each of the extracted sets of 2D body regions; and selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention will be better understood and readily apparent to one of ordinary skill in the art from the following written description, by way of example only, and in conjunction with the drawings, in which:

Figure 1 illustrates the block diagram of a human motion analysis system with the camera connected directly to the computer, according to an example embodiment.

Figure 2 shows a schematic top-down view drawing of an example embodiment comprising a camera.

Figure 3(a) illustrates the performer standing in a standard posture. Figure 3(b) illustrates a 3D model of the performer standing in a standard posture according to an example embodiment. The dots denote joints, straight lines denote bones connecting the joints, and gray scaled regions denote body parts.

Figure 4 illustrates an example of body region extraction. Figure 4(a) shows an input image and Figure 4(b) shows the extracted body regions, according to an example embodiment.

Figure 5 illustrates the flipping of the depth orientation of body part *b* in the *z*-direction to the new orientation denoted by a dashed line, according to an example embodiment.

Figure 6 illustrates an example result of posture candidate estimation according to an example embodiment Figure 6(a) shows the input image with a posture candidate overlaid. Figure 6(b) shows the skeletons of the posture candidates viewed from the front. At this viewing angle, all the posture candidates overlap exactly. Figure 6(c) shows

the skeletons of the posture candidates viewed from the side. Each candidate is shown with a different gray scale.

Figure 7 illustrates an example display of detailed 3D difference by overlapping the estimated performer's postures (dark gray scale) with the corresponding expert's postures (lighter gray scale) according to an example embodiment. The overlapping postures can be rotated in 3D to show different views. The estimated performer's postures can also be overlapped with the input images for visual verification of their correctness.

Figure 8 illustrates an example display of colored-coded regions overlapped with an input image for quick assessment according to an example embodiment. The darker gray scale regions indicate large error, the lighter gray scale regions indicate moderate error, and the transparent regions indicate negligible or no error.

Figure 9 illustrates the block diagram of a human motion analysis system with the camera and output device connected to the computer through a computer network, according to an example embodiment.

Figure 10 illustrates the block diagram of a human motion analysis system with the wireless input and output device, such as a hand phone or Personal Digital Assistant equipped with a camera, connected to the computer through a wireless network, according to an example embodiment.

Figure 11 shows a schematic top-down view of an example embodiment comprising multiple cameras arranged in a straight line.

Figure 12 shows a schematic top view of an example embodiment comprising multiple cameras placed around the performer.

Figure 13 shows a flow chart illustrating a method for human motion detection according to an example embodiment.

Figure 14 shows a schematic drawings of a computer system for implementing the method and system of an example embodiment.

DETAILED DESCRIPTION

The described example embodiments provide a system and method for acquiring a human performer's motion in one or more 2D videos, analyzing the 2D videos, comparing the performer's motion in the 2D videos and a 3D reference motion of an expert, computing the 3D differences between the performer's motion and the expert's

motion, and delivering information regarding the 3D difference to the performer for improving the performer's motion. The system in example embodiments comprises one or more 2D cameras, a computer, an external storage device, and a display device. In a single camera configuration, the camera acquires the performer's motion in a 2D video and passes the 2D video to a computing device. In a multiple camera configuration, the cameras acquire the performer's motion simultaneously in multiple 2D videos and pass the 2D videos to the computing device.

Some portions of the description which follows are explicitly or implicitly presented in terms of algorithms and functional or symbolic representations of operations on data within a computer memory. These algorithmic descriptions and functional or symbolic representations are the means used by those skilled in the data processing arts to convey most effectively the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities, such as electrical, magnetic or optical signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

Unless specifically stated otherwise, and as apparent from the following, it will be appreciated that throughout the present specification, discussions utilizing terms such as "calculating", "determining", "generating", "initializing", "outputting", or the like, refer to the action and processes of a computer system, or similar electronic device, that manipulates and transforms data represented as physical quantities within the the computer system into other data similarly represented as physical quantities within the computer system or other information storage, transmission or display devices.

The present specification also discloses apparatus for performing the operations of the methods. Such apparatus may be specially constructed for the required purposes, or may comprise a general purpose computer or other device selectively activated or reconfigured by a computer program stored in the computer. The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general purpose machines may be used with programs in accordance with the teachings herein. Alternatively, the construction of more specialized

apparatus to perform the required method steps may be appropriate. The structure of a conventional general purpose computer will appear from the description below.

In addition, the present specification also implicitly discloses a computer program, in that it would be apparent to the person skilled in the art that the individual steps of the method described herein may be put into effect by computer code. The computer program is not intended to be limited to any particular programming language and implementation thereof. It will be appreciated that a variety of programming languages and coding thereof may be used to implement the teachings of the disclosure contained herein. Moreover, the computer program is not intended to be limited to any particular control flow. There are many other variants of the computer program, which can use different control flows without departing from the spirit or scope of the invention.

Furthermore, one or more of the steps of the computer program may be performed in parallel rather than sequentially. Such a computer program may be stored on any computer readable medium. The computer readable medium may include storage devices such as magnetic or optical disks, memory chips, or other storage devices suitable for interfacing with a general purpose computer. The computer readable medium may also include a hard-wired medium such as exemplified in the Internet system, or wireless medium such as exemplified in the GSM mobile telephone system. The computer program when loaded and executed on such a general-purpose computer effectively results in an apparatus that implements the steps of the preferred method.

The invention may also be implemented as hardware modules. More particular, in the hardware sense, a module is a functional hardware unit designed for use with other components or modules. For example, a module may be implemented using discrete electronic components, or it can form a portion of an entire electronic circuit such as an Application Specific Integrated Circuit (ASIC). Numerous other possibilities exist. Those skilled in the art will appreciate that the system can also be implemented as a combination of hardware and software modules.

The motion analysis and comparison is performed in the following stages in an example embodiment:

1. Extracting the performer's body regions in each image frame of the 2D videos.

2. Calibrating the parameters of the cameras.

3. Estimating the temporal correspondence and rigid transformations that best align the postures in a 3D reference motion to the body regions in the image frames.

4. Estimating the 3D posture candidates that produce the human body regions in the image frames, using the results obtained in Stage 3 as the initial estimates. .

5. Selecting the 3D posture candidate that best matches the human body region in each time instant of the 2D video and refine the temporal correspondence between the 2D video and the 3D reference motion. In the case of multiple-camera configuration, the selected 3D posture candidate simultaneously best matches the human body regions in each time instant of the multiple 2D videos

6. Computing the 3D difference between the selected 3D posture candidates and the corresponding 3D reference posture. The 3D difference can include 3D joint angle difference, 3D velocity difference, etc. depending on the requirements of the application domain.

7. Visualizing and highlighting the 3D difference in a display device.

An example embodiment of the present invention provides a system and method for acquiring a human performer's motion in one 2D video, analyzing the 2D video, comparing the performer's motion in the 2D video and a 3D reference motion of an expert, computing the 3D differences between the performer's motion and the expert's motion, and delivering information regarding the 3D difference to the performer for improving the performer's motion.

Figure 1 shows a schematic block diagram of the example embodiment of a human motion analysis system 100. The system 100 comprises a camera unit 102 coupled to a processing unit, here in the form of a computer 104. The computer 104 is further coupled to an output device 106, and an external storage device 108.

With reference to Figure 2, the example embodiment comprises a stationary camera 200 with a fixed lens, which is used to acquire a 2D video *m'* of the performer's 202 entire motion. The 2D video is then analyzed and compared with a 3D reference motion *M* of an expert. The difference between the performer's 202 2D motion and the expert's 3D reference motion is computed. The system displays and highlights the difference in an output device 106 (Figure 1).

The software component implemented on the computer 104 (Figure 1) in the example embodiment comprises the following processing stages:

1. Extracting the input body region $S'_{t'}$ in each image $I'_{t'}$ at time t' of the video m'.

2. Calibrating the parameters of the camera 200.

3. Estimating the temporal correspondence $C(t')$ between input video time t' and reference time t and rigid transformations $T_{t'}$ that best align the posture $B_{C(t')}$ in the 3D reference motion to the body region $S'_{t'}$ in image $I'_{t'}$ for each time t'.

4. Estimating the 3D posture candidates $B'_{t''}$ that align with the input body regions $B'_{t'}$ in the input images $I'_{t'}$ using the results obtained in Stage 3 as the initial estimates.

5. Selecting the 3D posture candidate that best matches the input body region $S'_{t'}$ for each time t', and refine the temporal correspondence $C(t')$.

6. Computing the 3D difference between the selected 3D posture candidate $B'_{t''(t')}$ and the corresponding 3D reference posture $B_{C(t')}$ at each time t'.

7. Visualizing and highlighting the 3D difference in the display device 106 (Figure 1).

The method for Stage 1 in an example embodiment comprises a background subtraction technique described in [C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1998], an iterative graph-cut segmentation technique described in [C. Rother, V. Kolmogorov, and A. Blake. Grabcut – interactive foreground extraction using iterated graph cuts. In Proceedings of ACM SIGGRAPH, 2004], and a skin detection technique described in [M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. International Journal of Computer Vision, 46:81-96, 2002]. The contents of those references are hereby incorporated by cross references. In different example embodiments, for videos with simple background, background subtraction technique is sufficient. For videos with complex background, iterative graph-cut and skin detection techniques should be used. Figure 4 illustrates an example result of body region extraction. Figure 4(a) shows an input image and Figure 4(b) shows the extracted body region. The lighter gray scale region is extracted by the iterative graph-cut segmentation technique, and the darker gray scale parts are extracted using skin detection and iterative graph-cut segmentation techniques.

The method for Stage 2 in the example embodiment comprises computing the parameters of a scaled-orthographic camera projection, which include the camera's 3D rotation angle $(\theta_x, \theta_z, \theta_z)$, camera position (c_x, c_y), and scale factor s. It is assumed that the performer's posture at the first image frame of the video is the same as a standard calibration posture (for example, Figure 3). The method comprises the following steps:

1. Setting the camera parameters to default values: $\theta_x = \theta_y = \theta_z = 0$, $c_x = c_y = 0$, $s = 1$

2. Projecting a 3D model of the performer at calibration posture under the default camera parameters and render as a 2D projected body region. This step can be performed using OpenGL [OpenGL, www.opengl.org] in the example embodiment. The content of that reference is hereby incorporated by cross-reference. It is noted that in different example embodiments, the 3D model of the performer can be provided in different forms. For example, a template 3D model may be used, that has been generated to function as a generic template for a large cross section of possible performers. In another embodiment a 3D model of an actual performer may first be generated, which will involve an additional pre-processing step for generation of the customized 3D model, as will be appreciated and is understood by a person skilled in the art.

3. Computing the principal direction and the principal length h of the 2D projected model body region by applying principal component analysis (PCA) on the pixel positions in the projected model body region. The principal direction is the first eigenvector computed by PCA, and the principal length is the maximum length of the model body region along the principal direction.

4. Computing the principal direction and the principal length h' of the extracted captured body region in the first image frame of the video in a similar way.

5. Computing the camera scale $s = h' / h$.

6. Computing the camera position (c_x, c_y).

Compute the center (p'_x, p'_y) of the extracted body region and the center (p_x, p_y) of the 2D projected model body region.

Compute the camera position as the difference between the centers, i.e. $c_x=(p_x-p'_x) / s$ and $c_y = (p'_y - p_y) / s$.

7. Computing the camera rotation angle θ_z about Z-axis as the angular difference between the principal directions of the extracted body region and the 2D projected model body region. Camera rotation angles θ_x and θ_y are omitted.

The calibration method for stage 2 in the example embodiment thus derives the camera parameters for the particular human motion analysis system in question. It will be appreciated by a person skilled in the art that the same parameters can later be used for human motion analysis of a different performer, provided that the camera settings remain the same for the different performer. On the other hand, as mentioned above, a customized calibration using customized 3D models of an actual performer may be performed for each performer if desired, in different embodiments.

It is noted that in different embodiments, the method for stage S2 may comprise using other existing algorithms for the camera calibration, such as for example the "camera calibration tool box for MatLab" [www.vision.Caltech.edu/bouguetj/calib_doc/], the contents of which are hereby incorporated by cross-reference.

The method for Stage 3 in the example embodiment comprises estimating the approximate temporal correspondence $C(t')$ and the approximate rigid transformation $T_{t'}$ that best align the posture $B_{C(t')}$ in the 3D reference motion to the extracted body region $S_{t'}$ in image $I_{t'}$ for each time $t' = 0, \ldots, L'$, where $L'+1$ is the length of the video sequence. The length of the 3D reference motion is $L+1$, for $t = 0, \ldots, L$. The estimation is subjected to a temporal order constraint: for any two temporally ordered postures in the performer's motion, the two corresponding postures in the reference motion have the same temporal order. That is, for any t_1 and t_2, such that $t_1 < t_2$, $C(t_1) < C(t'_2)$.

Given a particular C, each transformation $T_{t'}$ at time t' can be determined by finding the best match between extracted body region $S'_{t'}$ and 2D projected model body region $P(T_{t'}(BC(t')))$:

$$T_{t'} = \arg\min_{T} d_S(P(T(B_{C(t')})), S'_{t'})$$

where the optimal $T_{t'}$ is computed using a sampling technique described in [Sampling methods, www.statpac.com/surveys/sampling/htm]. The content of that reference is hereby incorporated by cross reference.

The method for computing the difference $d_S(S, S')$ between two image regions S and S' comprises computing two parts:

$$d_S(S, S') = \lambda_A d_A(A, A') + \lambda_E d_E(E, E')$$

where d_A is the amount of overlap between the set A of pixels in the silhouette of the 2D projected model body region and the set and A' of pixels in the silhouette of the extracted body region in the video image, d_E is the Chamfer distance described in [M.A. Butt and P. Maragos, Optimum design of chamfer distance transforms, IEEE Transactions on Image Processing, 7(10), 1998, 1477-1484] between the set E of edges in the 2D projected model body region and the set E' of edges in the extracted body region, and λ_A and λ_E are constant parameters. The content of that reference is hereby incorporated by cross-reference.

The method of computing the optimal temporal correspondence $C(t')$ comprises the application of dynamic programming as follows. Let $d(t', C(t'))$ denote the difference d_s:

$$d(t', C(t')) = d_S(P(T_{t'}(B_{C(t')})), S'_{t'})$$

Let **D** denote a $(L' + 1) \times (L + 1)$ correspondence matrix. Each matrix element at (t', t) corresponds to the possible frame correspondence between t' and t, and the correspondence cost is $d(t', t)$. A path in **D** is a sequence of frame correspondences for $t' = 0, \ldots, L'$ such that each t' has a unique corresponding $t = C(t')$. It is assumed that $C(0) = 0$ and $C(L') = L$. Let $D(t', t)$ denote the least cost from the frame pair (0, 0) up to (t', t) on the least cost path, and $D(0, 0) = d(0, 0)$. Then, the optimal solution given by $D(L', L)$ can be recursively computed using dynamic programming as follows:

$$D(t', t) = d(t', t) + \min_{i=0}^{w} D(t' - 1, t - 1 - i)$$

Once $D(L', L)$ is computed, the least cost path is obtained by tracing back the path from $D(L', L)$ to $D(0, 0)$. The optimal $C(t')$ is given by the least cost path.

The method for stage 4 in the example embodiment estimates 3D posture candidates that align with the extracted body regions. That is, for each time t', find a set $\{B'_{t'}\}$ of 3D posture candidates whose 2D projected model body regions $P(A_{t'}(T_{t'}(B'_{t'})))$ match the extracted body region $B'_{t'}$ in the input images $I'_{t'}$. The computation of the 3D posture candidates is subjected to the joint angle limit constraint: The valid joint rotation of each body part is limited to physically possible ranges.

The example embodiment uses a nonparametric implementation of the Belief Propagation (BP) technique described in [E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky. Nonparametric belief propagation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 605-612, 2003. M. Isard. Pampas: Real-valued graphical models for computer vision. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 613-620, 2003. G. Hua and Y. Wu. Multi-scale visual tracking by sequential belief propagation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 826-833, 2004. E.B. Sudderth, M.I. Mandel, W.T. Freeman, and A.S. Willsky. Visual hand tracking using nonparametric belief propagation. In IEEE CVPR Workshop on Generative Model based Vision, 2004.]. The contents of those references are hereby incorporated by cross reference.

It comprises the following steps:

1. Run the nonparametric BP algorithm to generate pose samples for each body part using the results in Stage 3 as the initial estimates. That is, based on the results in Stage 3, the temporary align posture in the 3D reference motion forms the initial estimate for each frame.

2. Determine a best matching pose for each body part.

• If the pose samples of each body part converge to a single state, choose any pose sample as the best pose for this body part.

• If the pose samples of each body part do not converge to a single state, project each body part at each pose sample to compute the mean image positions of its joints. Then, starting from the root body part, generate a pose sample for each body part such that the body part at the pose sample is connected to its parent body part, and the projected image positions of its joints match the computed mean positions of its joints.

3. Generate the first posture candidate. For each body parts starting from the root body part, modify the depth orientation of the best pose sample such that it has the same depth orientation as that in the corresponding reference posture. All the pose samples are combined into a posture candidate by translating the depth coordinate in each sample, if necessary, such that the neighboring body parts are connected.

4. Generate new 3D posture candidates. Starting from the first 3D posture candidate, flip the depth orientation of n body parts about their parent joints, starting with $n = 1$, while keeping the body parts connected at the joints. Figure 5 illustrates flipping of body part b from a position k' to k, around a parent joint at j.

5. The above step is repeated for n = 1, 2, . . . , until N posture candidates are generated.

Figure 6 illustrates example posture candidates in Figures 6(b) and (c) generated from an input image in Figure 6(a). In Figure 6(b) the skeletons of the posture candidates are viewed from the front. At this viewing angle, all the posture candidates overlap exactly, given the nature of how they have been derived explained above for the example embodiment. Figure 6(c) shows the different skeletons of the posture candidates viewed from the side, illustrating the differences between the respective posture candidates.

The method for Stage 5 in the example embodiment comprises refining the estimate of temporal correspondence $C(t')$ and selecting the best posture candidates $B'_{t'l'}$ that best match the corresponding reference postures $B_{C(t')}$.

The refinement is subjected to temporal ordering constraint: for any t'^1 and t'^2, such that $t'^1 < t'^2$, $C(t'^1) < C(t'^2)$, and a constraint of small rate of change of posture errors: for each t', $\Delta\varepsilon_{t'} / \Delta t' = (\varepsilon_{t'} - \varepsilon_{t'-\Delta t'}) / \Delta t'$ is small.

The method of computing the optimal refined temporal correspondence $C(t')$ comprises the application of dynamic programming as follows. Let $d_c(t', t, l')$ denote the 3D posture difference between the posture candidate $B'_{t'l'}$ and the reference posture B_t, which is measured as the mean difference between the orientations of the bones in the postures. Let $d_s(t', t, s, l', k')$ denote the change of posture difference between the corresponding pairs $(B'_{t'l'}, B_t)$ and $(B'_{t'-1,k'}, B_s)$.

Let **D** denote a $(L' + 1) \times (L + 1) \times N$ correspondence matrix, where N is the maximum number of posture candidates at any time t'. Each matrix element at (t', t, l') corresponds to the possible correspondence between t', t, and l', and the correspondence cost is $d(t', t, l')$. A path in **D** is a sequence of correspondences for $t' = 0, \ldots, L'$ such that each t' has a unique corresponding $t = C(t')$ and a unique corresponding posture candidate $l' = l(t')$. It is assumed that $C(0) = 0$ and $C(L') = L$. Let $D(t', t, l')$ denote the least cost from the triplet $(0, 0, l'_0)$ up to (t', t, l') on the least cost path, and $D(0, 0, l'_0) = d_c(0, 0, l'_0)$. Then, the optimal solution given by $D(L', L, l(L'))$ can be recursively computed using dynamic programming as follows:

$$D(t', t, \ell(t')) = \min_{l'} D(t', t, l')$$

$$\ell(t') = \arg\min_{l'} D(t', t, l')$$

where

$$D(t', t, l') = d_c(t', t, l') + \min_{i,k'}\{D(t'-1, t-1-i, k') + d_s(t', t, t-1-i, l', k')\}$$

Once $D(L', L, l(L'))$ is computed, the least cost path is obtained by tracing back the path from $D(L', L, l(L'))$ to $D(0, 0, l(0))$. The optimal $C(t')$ and $l(t')$ are given by the least cost path.

The method for Stage 6 in the example embodiment comprises computing the 3D difference between the selected 3D posture candidate $B'_{t'l(t')}$ and the corresponding 3D reference posture $B_{C(t')}$ at each time t'. The 3D difference can include 3D joint angle difference, 3D joint velocity difference, etc. depending on the specific coaching requirements of the sports.

The method for Stage 7 in the example embodiment comprises displaying and highlighting the 3D difference in a display device. An example display of detailed 3D difference is illustrated in Figure 7. Figure 7 illustrates an example display of detailed 3D difference by overlapping the estimated performer's postures e.g. 700 (dark gray scale) with the corresponding expert's postures e.g. 702 (lighter gray scale) according to an example embodiment. The overlapping postures can be rotated in 3D to show different views (compare rows 704 and 706). The estimated performer's postures can also be overlapped with the input images (row 708) for visual verification of their correctness.

An example display of color-coded errors for quick assessment is illustrated in Figure 8. Figure 8 illustrates an example display of colored-coded regions e.g. 800, 802 overlapped with an input image 804 for quick assessment according to an example embodiment. The darker gray scale regions e.g. 800 indicates large error, the lighter gray scale regions e.g. 802 indicates moderate error, and the transparent regions e.g. 806 indicate negligible or no error.

In another embodiment where the 3D reference motion contains multiple pre-defined motion segments, such as Taichi motion, the 2D input video is first segmented

into the corresponding performer's motion segments. The method of determining the corresponding performer's segment boundary for each reference segment boundary t, comprises the following steps:

1. Determine initial estimate of the performer's motion segment boundary t' by $C(t') = t$.

2. Obtain a temporal window $[t' - \omega, t' + \omega]$, where ω is the window size.

3. Find one or more smooth sequences of posture candidates in the temporal window.

• Correct posture candidates should change smoothly over time. Suppose $B'_{\tau,l'}$ and $B'_{\tau+1,k'}$ are correct posture candidates, then the 3D posture difference between them $d_B(B'_{\tau,l'}, B'_{\tau+1,k'})$, which is measured as the mean difference between the orientations of the bones in the postures, is small for any $\tau \in [t' - \omega, t' + \omega]$.

• Choose a posture candidate for each $\tau \in [t' - \omega, t' + \omega]$ to obtain a sequence of posture candidates that satisfy the condition that $d_B(B'_{\tau,l'}, B'_{\tau+1,k'})$ is small for each τ.

4. Find candidate segment boundaries.

• For each smooth sequence of posture candidates, find the candidate segment boundary $\tau \in [t' - \omega, t' + \omega]$ and the corresponding posture candidate at τ that satisfies the segment boundary condition: At a segment boundary, there are large changes of motion directions for some joints.

• Denote a candidate segment boundary found above as τ_i and the corresponding posture candidate as $B'_{i'}$.

5. Identify the optimal segment boundary τ^*.

The posture candidate at the optimal segment boundary τ^* should be the most similar to the corresponding reference posture B_t. Therefore, τ can be determined as follows,

$$k = \arg\min_i d_B(B_t, B'_i),$$

$$\tau^* = \tau_k .$$

In another example embodiment, the input body region is extracted with the help of colored markers.

In another example embodiment, the appendages carried by the performer, e.g., a golf club, is also segmented.

In another example embodiment, the 3D reference motion of the expert is replaced by the 3D posture sequence of the performer computed from the input video acquired in a previous session.

In another example embodiment, the 3D reference motion of the expert is replaced by the 3D posture sequence of the performer computed from the input videos acquired in previous sessions that best matches the 3D reference motion of the expert.

In another example embodiment, the camera 900 and output device 902 are connected to a computer 904 through a computer network 906, as shown in Figure 9. The computer 904 is coupled to an external storage device 908 directly in this example.

In another example embodiment, a wireless input and output device 1000, such as a hand phone or Personal Digital Assistant equipped with a camera, is connected to a computer 1002 through a wireless network 1004, as shown in Figure 10. The computer 1002 is coupled to an external storage device 1006 directly in this example.

In another example embodiment, multiple cameras 1101-1103 are arranged along a straight line, as shown in Figure 11. Each camera acquires a portion of the performer's 1104 entire motion when the performer 1104 passes in front of the respective camera. This embodiment also allows the system to acquire high-resolution video of a user whose body motion spans a large arena.

In another example embodiment, multiple cameras 1201-1204 are placed around the performer 1206, as shown in Figure 12. This arrangement allows different cameras to capture the frontal view of the performer 1206 when he faces different cameras.

In another example embodiment, the arrangements of the cameras discussed above are combined.

In the multi-camera configurations in different example embodiments, for example those shown in Figures 11 and 12, the calibration method for the stage S2 processing, in addition to calibration of each of the individual cameras as described

above for the single camera embodiment, further comprises computing the relative positions and orientations between the cameras using an inter-relation algorithm between the cameras, as will be appreciated by a person skilled in the art. Such inter-relation algorithms are understood in the art, and will not be described in more detail herein. Reference is made for example to [R. Jain, R. Kasturi, and B. G. Schunck, Machine Vision, McGraw-Hill 1995] and [R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.] for example algorithms for use in such an embodiment. The contents of those references are hereby incorporated by cross-reference.

Example embodiments of the method and system for human motion analysis can have the following framework of stages:

1. Input Video Segmentation

This stage segments the human body in each image frame of the input video. The human body, the arms, and the background are assumed to have different colors so that they can be separated. This assumption is reasonable and easily satisfied, for instance, for a user who wears short-sleeved colored shirt and stands in front of a background of a different color. The background can be a natural scene which is nonuniform in color. This stage is achieved using a combination of background removal, graph-cut algorithm and skin color detection. In case the background is uniform, the segmentation algorithm can be simplified.

2. Camera Calibration

This stage computes the camera's extrinsic parameters, assuming that its intrinsic parameters have already been pre-computed. This stage can be achieved using existing camera calibration algorithms.

3. Estimation of Approximate Temporal Correspondence

This stage estimates the approximate temporal correspondence between 3D reference motion and 2D input video. Dynamic Programming technique is used to estimate the temporal correspondence between the input video and the reference motion by matching the 2D projections of 3D postures in the reference motion with the segmented human body in the 2D input video. This stage also estimates the approximate global rotation and translation of the user's body relative to the 3D reference motion.

4. Estimation of Posture Candidates

This stage estimates, for each 2D input video frame, a set of 3D posture candidates that can produce 2D projections that are the same as that in the input video frame. This is performed using an improved version of Belief Propagation method. In a single-camera system, these sets typically have more than one posture candidate each due to depth ambiguity and occlusion. In a multiple-camera system, the number of posture candidates may be reduced.

5. Selection of best posture candidates

This stage selects the best posture candidates that form smooth motion over time. It also refines the temporal correspondence estimated in Stage 2. This stage is accomplished using Dynamic Programming.

The framework of the example embodiments can be applied to analyze various types of motion by adopting appropriate 3D reference motion. It will be appreciated by a person skilled in the art that by adapting the system and method to handle specific application domains, these stages can be refined and optimized to reduce computational costs and improve efficiency.

Figure 13 shows a flow chart 1300 illustrating a method for human motion detection according to an example embodiment. At step 1302, one or more 2D input videos of the human motion are captured. At step 1304, sets of 2D body regions are extracted from respective frames of the 2D input videos. At step 1306, 3D human posture candidates are determined for each of the extracted sets of 2D body regions. At step 1308, a sequence of 3D human postures from the 3D human posture candidates for the respective frames is selected as representing the human motion in 3D.

The method and system of the example embodiment can be implemented on a computer system 1400, schematically shown in Figure 14. It may be implemented as software, such as a computer program being executed within the computer system 1400, and instructing the computer system 1400 to conduct the method of the example embodiment.

The computer system 1400 comprises a computer module 1402, input modules such as a keyboard 1404 and mouse 1406 and a plurality of output devices such as a display 1408, and printer 1410.

The computer module 1402 is connected to a computer network 1412 via a suitable transceiver device 1414, to enable access to e.g. the Internet or other network systems such as Local Area Network (LAN) or Wide Area Network (WAN).

The computer module 1402 in the example includes a processor 1418, a Random Access Memory (RAM) 1420 and a Read Only Memory (ROM) 1422. The computer module 1402 also includes a number of Input/Output (I/O) interfaces, for example I/O interface 1424 to the display 1408, and I/O interface 1426 to the keyboard 1404.

The components of the computer module 1402 typically communicate via an interconnected bus 1428 and in a manner known to the person skilled in the relevant art.

The application program is typically supplied to the user of the computer system 1400 encoded on a data storage medium such as a CD-ROM or flash memory carrier and read utilising a corresponding data storage medium drive of a data storage device 1430. The application program is read and controlled in its execution by the processor 1418. Intermediate storage of program data maybe accomplished using RAM 1420.

It will be appreciated by a person skilled in the art that numerous variations and/or modifications may be made to the present invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects to be illustrative and not restrictive.

CLAIMS

1. A method for human motion analysis, the method comprising the steps of:

capturing one or more 2D input videos of the human motion;

extracting sets of 2D body regions from respective frames of the 2D input videos;

determining 3D human posture candidates for each of the extracted sets of 2D body regions; and

selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

2. The method as claimed in claim 1, further comprising the step of determining differences between 3D reference data for said human motion and the selected sequence of 3D human postures.

3. The method as claimed in claim 2, further comprising the step of visualizing said differences to a user.

4. The method as claimed in any one of the preceding claims, wherein extracting the sets of 2D body regions comprises one or more of a group consisting of background subtraction, iterative graph-cut segmentation and skin detection.

5. The method as claimed in any one of the preceding claims, wherein determining the 3D human posture candidates comprises the steps of:

generating a first 3D human posture candidate; and

flipping a depth orientation of body parts represented in the first 3D human posture candidate around respective joints to generate further 3D human posture candidates from the first 3D human posture candidate.

6. The method as claimed in claim 5, wherein generating the first 3D human posture candidate comprises temporally aligning the extracted sets of 2D body portions from each frame with 3D reference data of the human motion and adjusting the 3D reference data to match the 2D body portions.

7. The method as claimed in any one of the preceding claims, wherein selecting the sequence of 3D human postures from the 3D human posture candidates is based on a least cost path among the 3D human posture candidates for the respective frames.

8. The method as claimed in claim 7, wherein selecting the sequence of 3D human postures from the 3D human posture candidates further comprises refining a temporal alignment of the extracted sets of 2D body portions from each frame with 3D reference data of the human motion.

9. A system for human motion analysis, the method comprising the steps of:

means for capturing one or more 2D input videos of the human motion;

means for extracting sets of 2D body regions from respective frames of the 2D input videos;

means for determining 3D human posture candidates for each of the extracted sets of 2D body regions; and

means for selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

10. The system as claimed in claim 9, further comprising means for determining differences between 3D reference data for said human motion and the selected sequence of 3D human postures.

11. The system as claimed in claim 10, further comprising means for visualizing said differences to a user.

12. The system as claimed in any one of claims 9 to 11, wherein the means for extracting the sets of 2D body regions performs one or more of a group consisting of background subtraction, iterative graph-cut segmentation and skin detection.

13. The system as claimed in any one of claims 9 to 12, wherein the means for determining the 3D human posture candidates generates a first 3D human posture candidate; and flips a depth orientation of body parts represented in the first 3D human posture candidate around respective joints to generate further 3D human posture candidates from the first 3D human posture candidate.

14. The system as claimed in claim 13, wherein generating the first 3D human posture candidate comprises temporally aligning the extracted sets of 2D body portions from each frame with 3D reference data of the human motion and adjusting the 3D reference data to match the 2D body portions.

15. The system as claimed in any one of claims 9 to 14, wherein the means for selecting the sequence of 3D human postures from the 3D human posture candidates determines a least cost path among the 3D human posture candidates for the respective frames.

16. The system as claimed in claim 15, wherein the means for selecting the sequence of 3D human postures from the 3D human posture candidates further comprises means for refining a temporal alignment of the extracted sets of 2D body portions from each frame with 3D reference data of the human motion.

17. A data storage medium having computer code means for instructing a computing device to execute a method for human motion detection, the method comprising the steps of:

capturing one or more 2D input videos of the human motion;

extracting sets of 2D body regions from respective frames of the 2D input videos;

determining 3D human posture candidates for each of the extracted sets of 2D body regions; and

selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D.

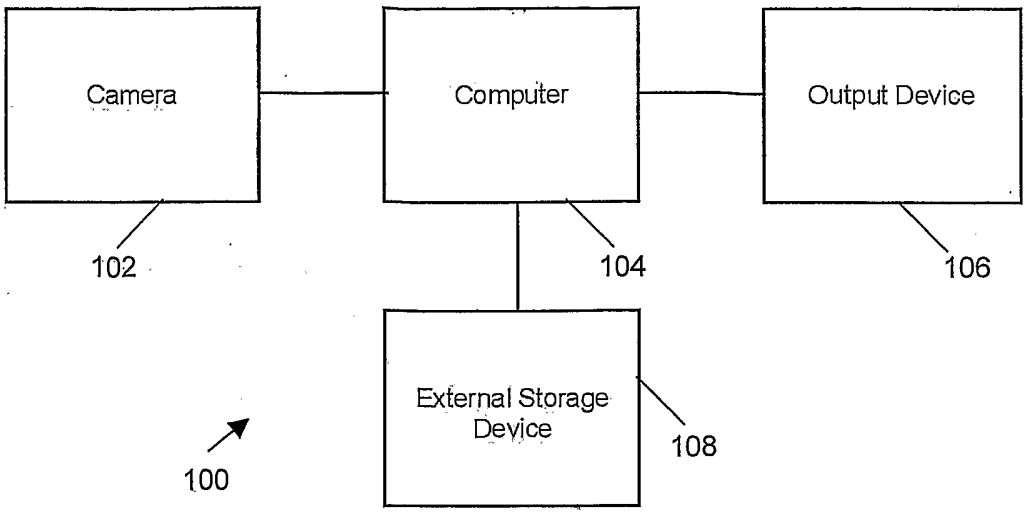

Figure 1

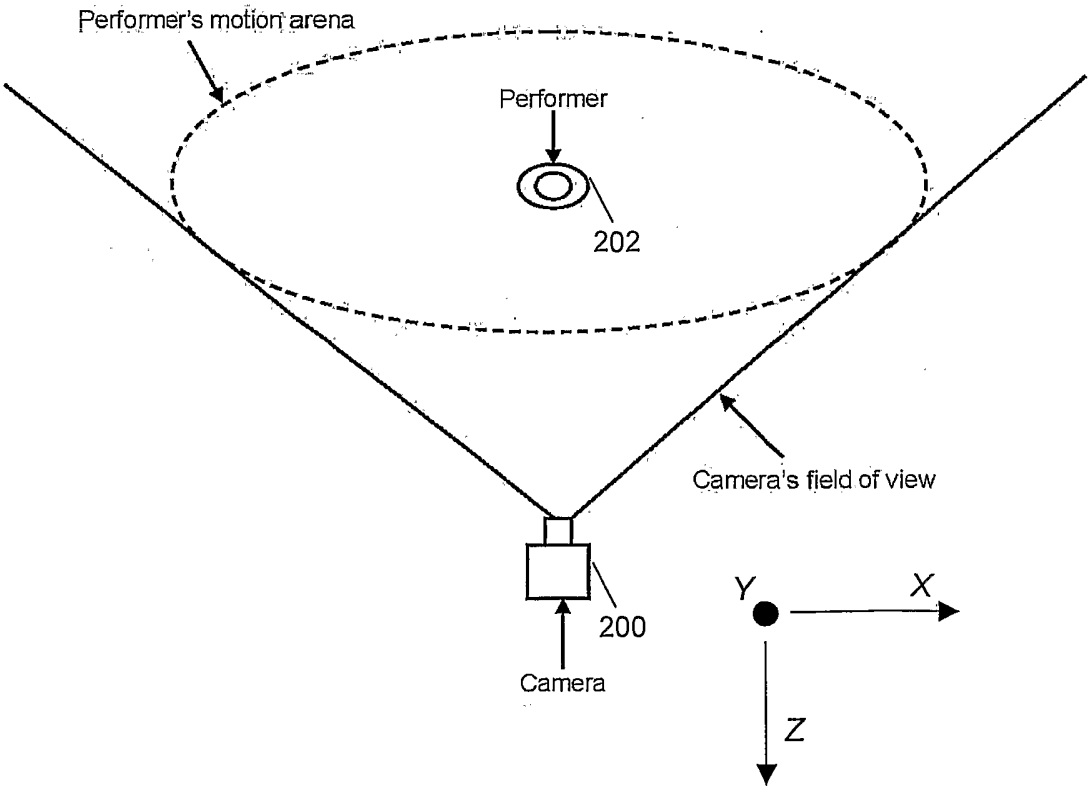

Figure 2

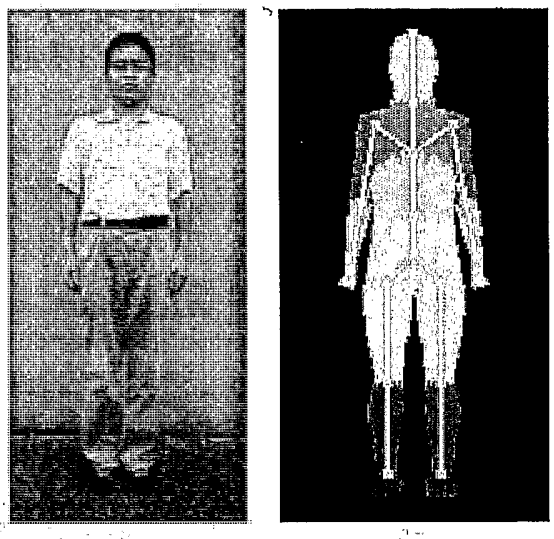

Figure 3

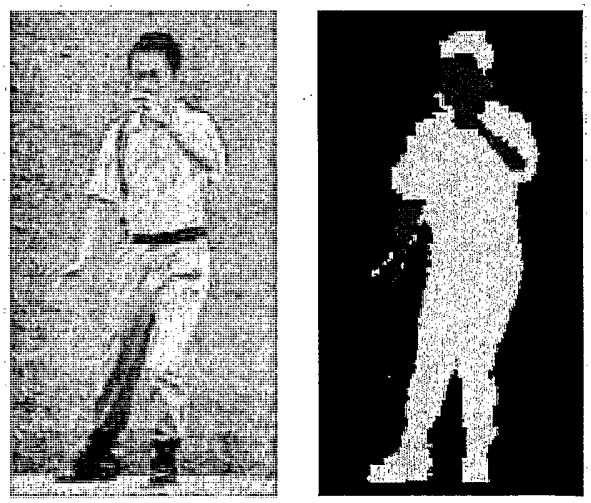

Figure 4

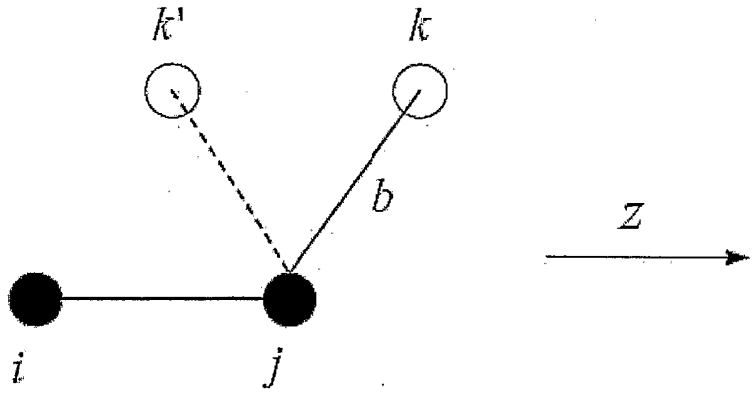

Figure 5

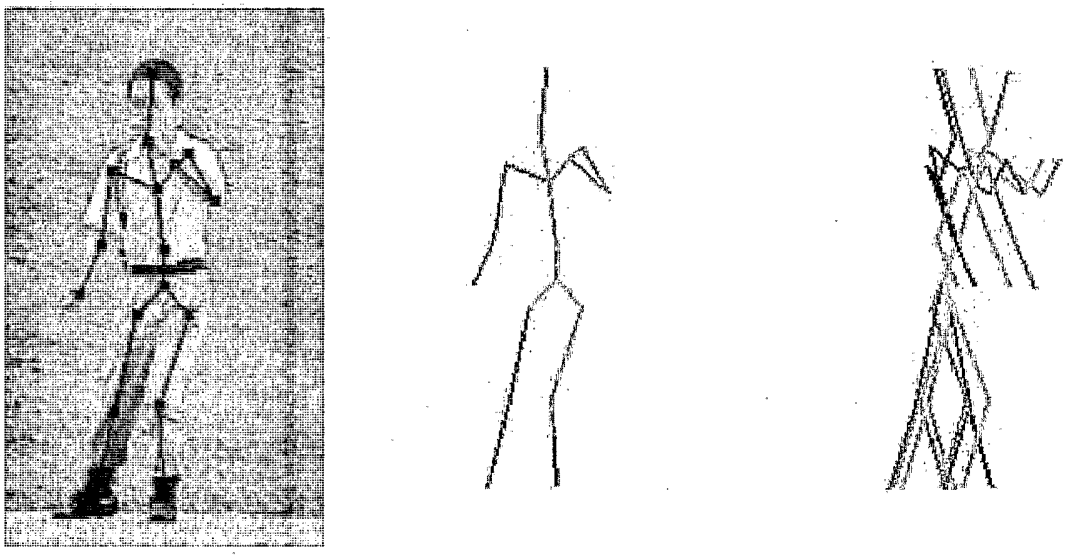

Figure 6

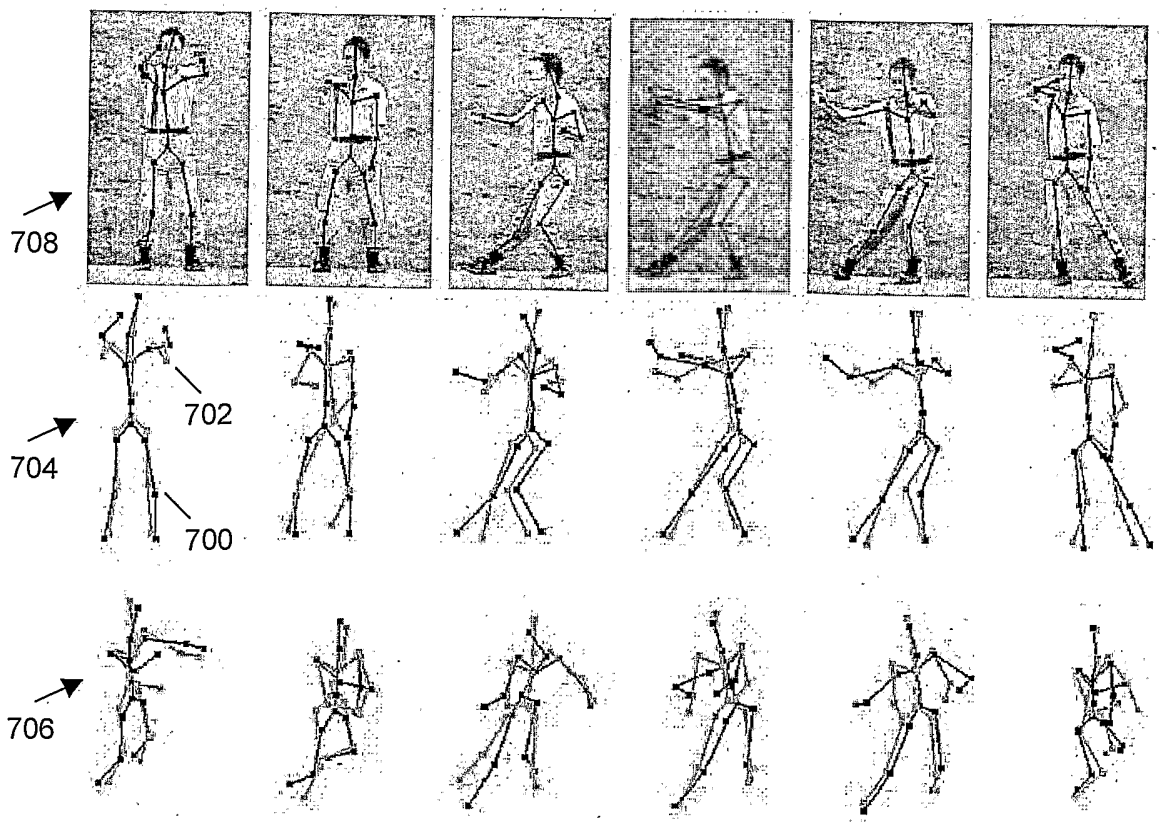

Figure 7

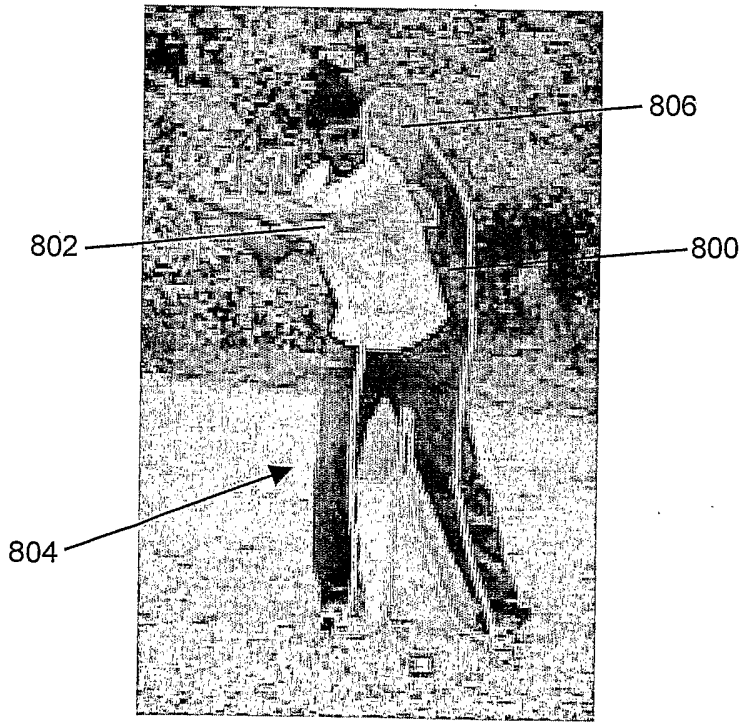

Figure 8

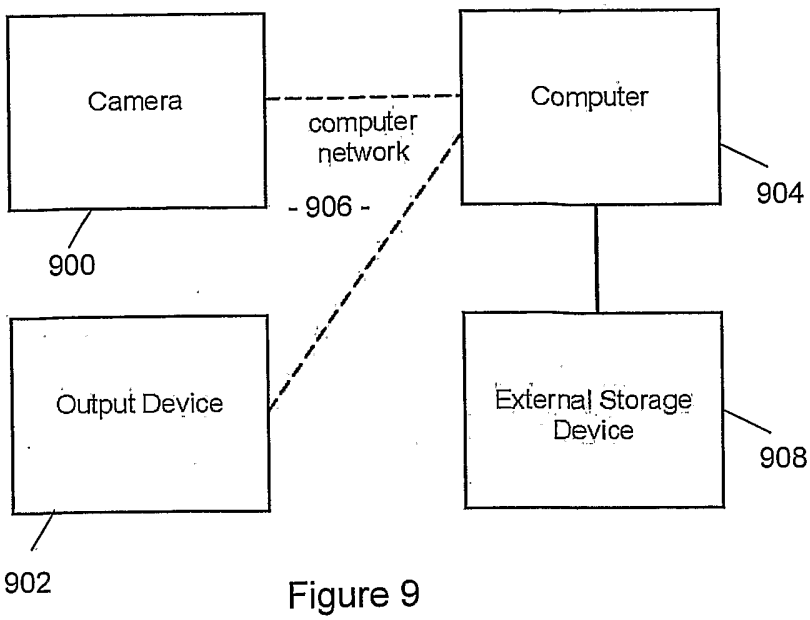

Figure 9

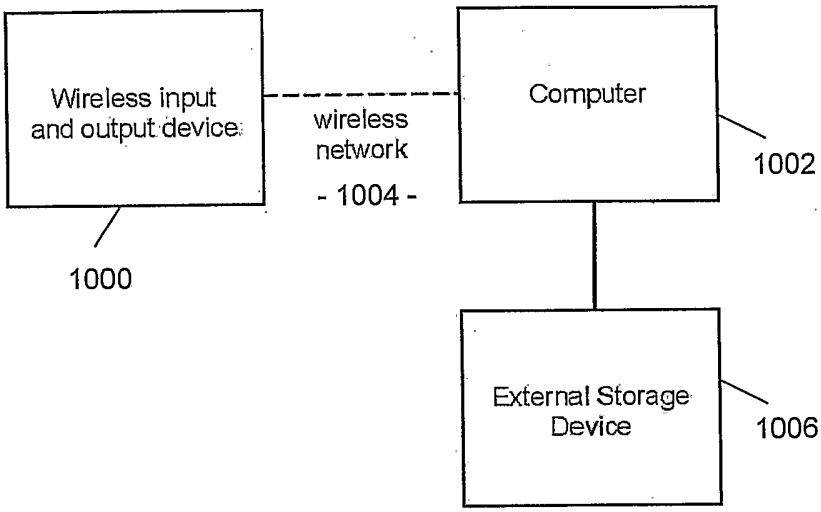

Figure 10

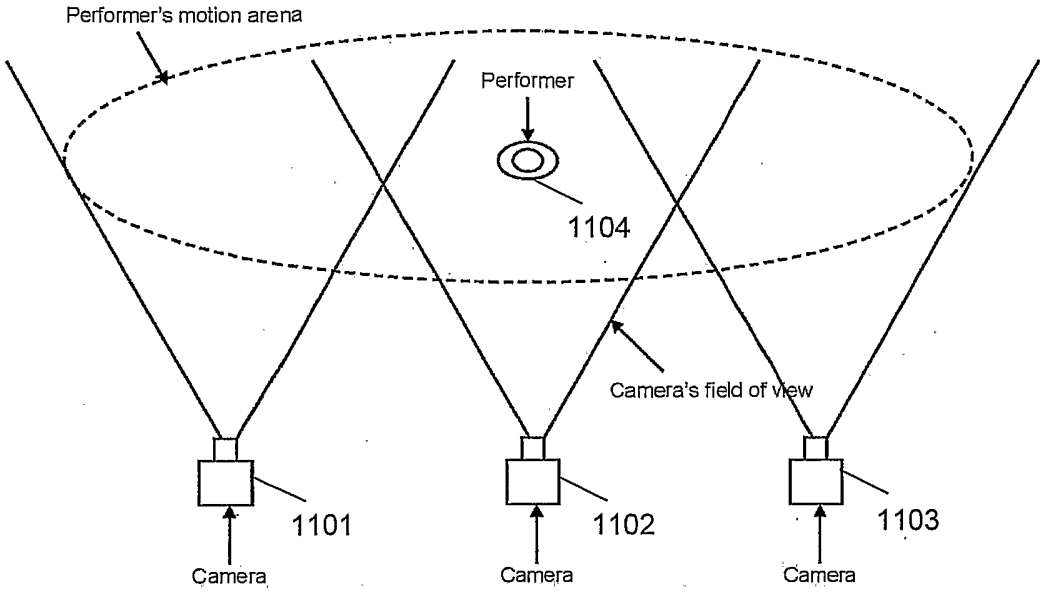

Figure 11

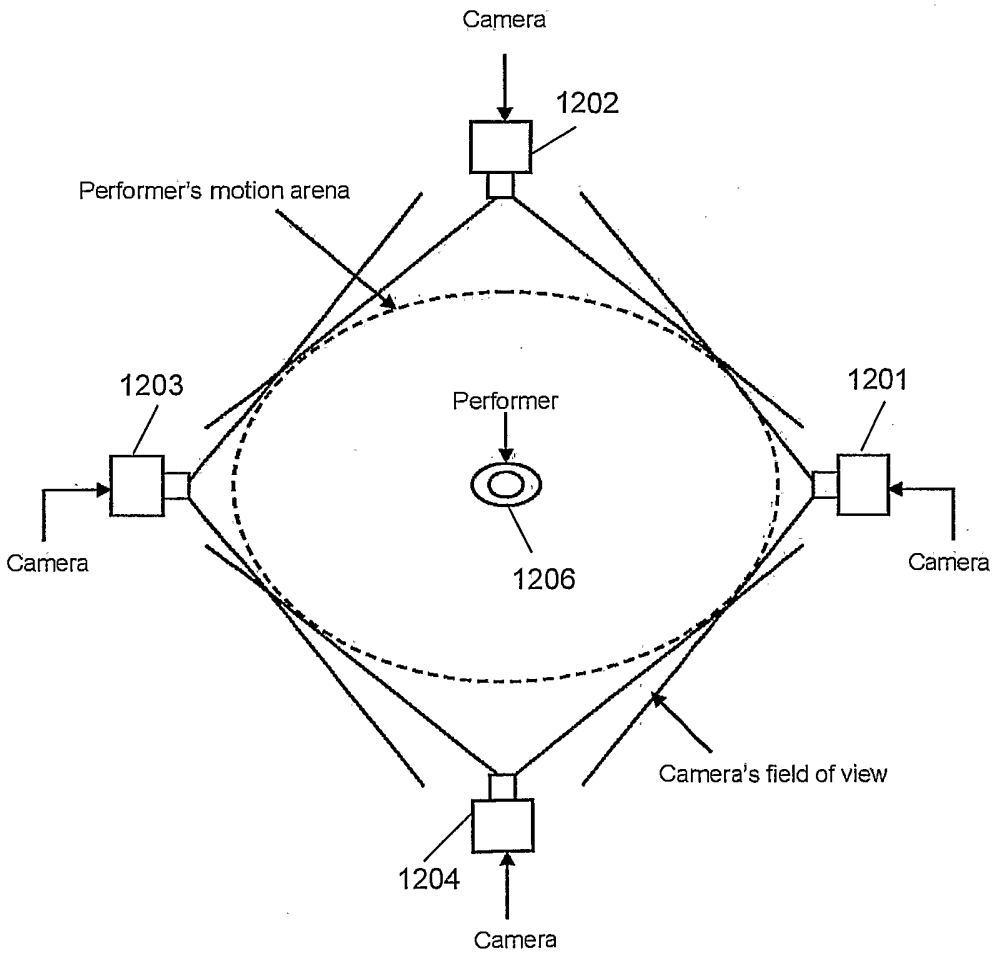

Figure 12

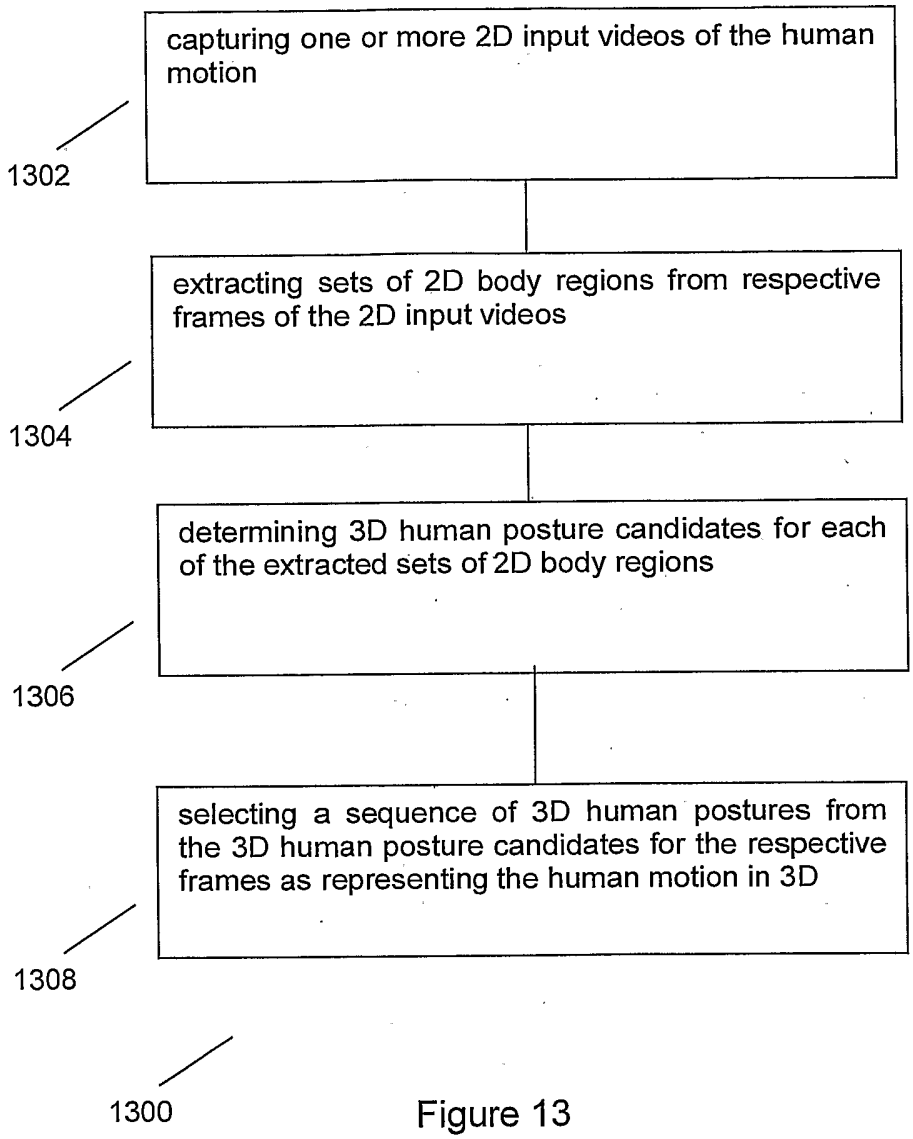
capturing one or more 2D input videos of the human motion
1302
extracting sets of 2D body regions from respective frames of the 2D input videos
1304
determining 3D human posture candidates for each of the extracted sets of 2D body regions
1306
selecting a sequence of 3D human postures from the 3D human posture candidates for the respective frames as representing the human motion in 3D
1308
1300

Figure 13

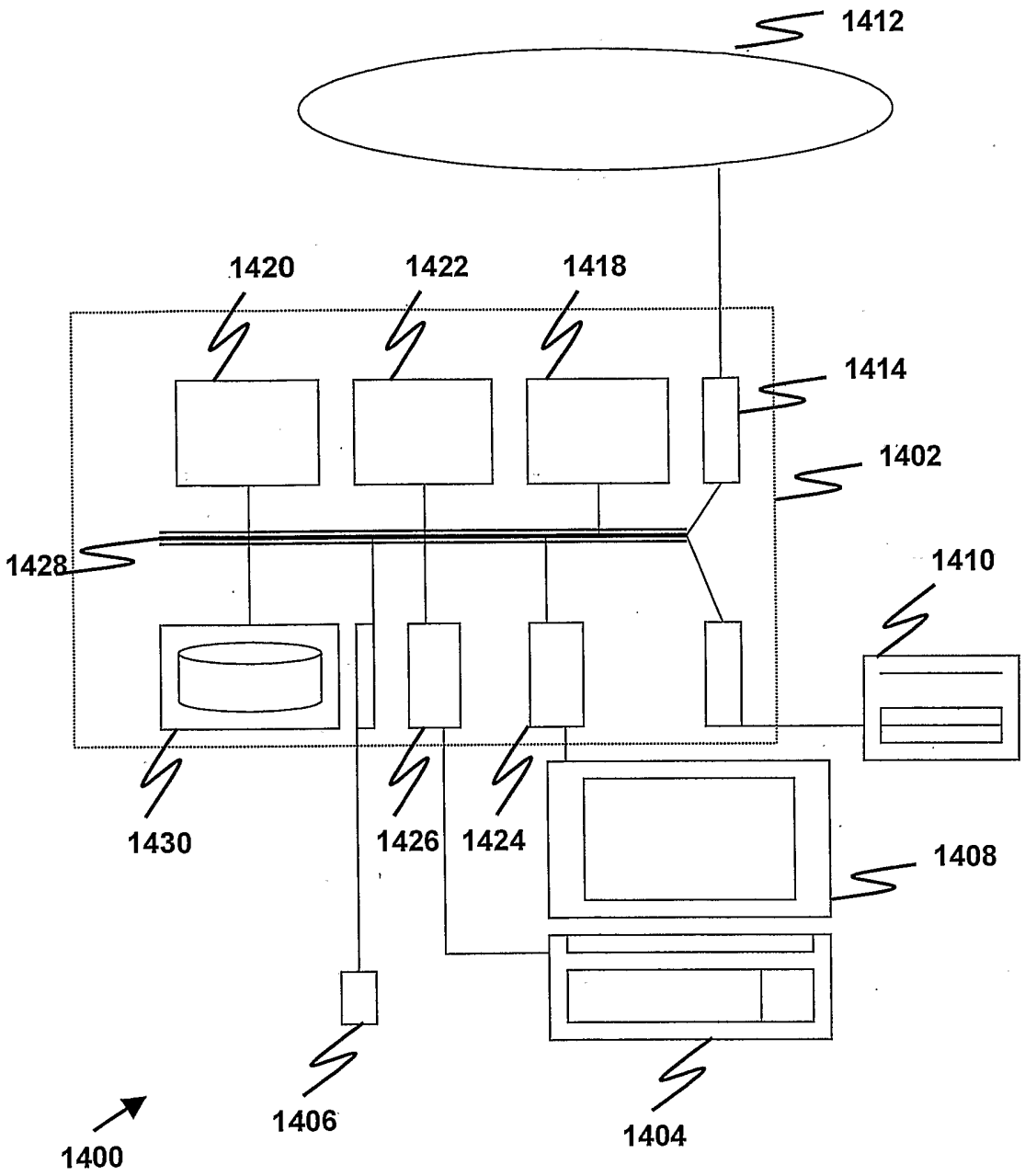
1412
1420
1422
1418
1414
1402
1428
1410
1430
1426
1424
1408
1406
1404
1400

Figure 14