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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 March 2016 (03.03.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/033279 A1

(51) International Patent Classification:
G06F 3/01 (2006.01)

(21) International Application Number:
PCT/US2015/047095

(22) International Filing Date:
27 August 2015 (27.08.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14/473,679 29 August 2014 (29.08.2014) US

(71) Applicant: **KONICA MINOLTA LABORATORY U.S.A., INC.** [US/US]; 2855 Campus Drive, Suite 100, San Mateo, California 94403 (US).

(72) Inventors: **AUGE, Quentin**; c/o Konica Minolta Laboratory U.S.A., Inc., 2855 Campus Drive, Suite 100, San Mateo, California 94403 (US). **ZHANG, Yongmian**; 33001 Calistoga Street, Union City, California 94587 (US). **GU, Haisong**; 22861 Medina Lane, Cupertino, California 95014 (US).

(74) Agent: **NUZUM, Kirk M.**; Buchanan Ingersoll & Rooney PC, P.O. Box 1404, Alexandria, Virginia 22313-1404 (US).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

(54) Title: METHOD AND SYSTEM OF TEMPORAL SEGMENTATION FOR GESTURE ANALYSIS

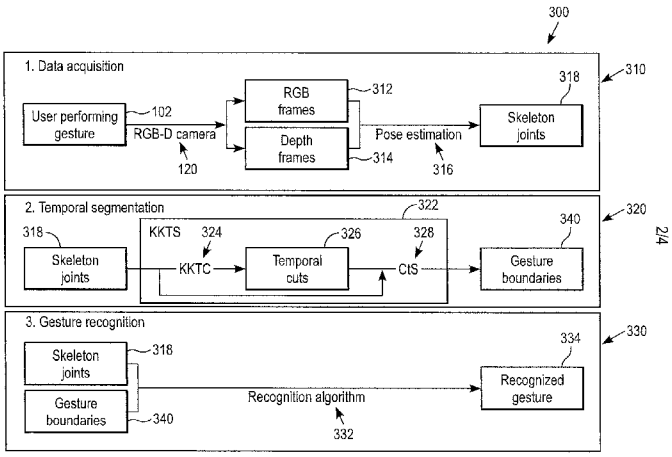


FIG. 3

(57) Abstract: A method, system and non-transitory computer readable medium are disclosed for recognizing gestures, the method includes capturing at least one three-dimensional (3D) video stream of data on a subject; extracting a time-series of skeletal data from the at least one 3D video stream of data; isolating a plurality of points of abrupt content change called temporal cuts, the plurality of temporal cuts defining a set of non-overlapping adjacent segments partitioning the time-series of skeletal data; identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration; and classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

WO 2016/033279 A1

METHOD AND SYSTEM OF TEMPORAL SEGMENTATION FOR GESTURE ANALYSIS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Patent Application Serial No. 14/473,679, filed August 29, 2014, the entire contents of which is hereby incorporated herein by reference.

FIELD

[0002] The present disclosure relates to a method and system of temporal segmentation for gesture analysis, and more particularly, to a method and system for identifying gesture boundaries within a flow of human motion, which can be used as input or preprocessing module for gesture analysis, such as gesture classification and recognition.

BACKGROUND

[0003] Gesture recognition is an example of an application using an efficient temporal segmentation, or the task of finding gestures within a flow of human motion, as a pre-processing step. Usually performed in an unsupervised manner, the step of temporal segmentation facilitates subsequent recognition of gestures.

[0004] Gesture recognition and segmentation can be performed either in a simultaneous or sequential fashion. For example, machine learning frameworks capable of modeling time aspects directly, such as hidden Markov models (HMMs), continuous-time recurrent neural networks (CTRNNs), dynamic Bayesian network (DBNs) or conditional random fields (CRFs) can be used for simultaneous gesture recognition and segmentation. Temporal segmentation has also been studied independently of its recognition counterpart. Nevertheless, when it occurs, two main approaches predominate, namely temporal clustering and change-point detection.

[0005] Temporal clustering (TC) refers to the factorization of multiple time series into a set on non-overlapping segments that belongs to k temporal clusters. Being inherently offline, the approach benefits from a global point of view on the data and provides cluster labels as in clustering. However, temporal clustering may not be suitable for real-time applications.

[0006] Change-point methods rely on various tools from signal theory and statistics to locate frames of abrupt change in pattern within the flow of motion. Although change-point methods can be restricted to univariate series with parametric

distribution assumption (which does not hold when analyzing human motion), the recent use of kernel methods released part of these limitations, change-point methods have been recently applied to the temporal segmentation problem. Unlike temporal clustering, the change-point approach often results in unsupervised online algorithms, which can perform real-time, relying on local patterns in time-series.

[0007] Although significant progress has been made in temporal segmentation, the problem still remains inherently challenging due to viewpoint changes, partial occlusions, and spatio-temporal variations.

SUMMARY

[0008] In accordance with an exemplary embodiment, a method is disclosed for recognizing gestures, comprising: capturing at least one three-dimensional (3D) video stream of data on a subject; extracting a time-series of skeletal data from the at least one 3D video stream of data; isolating a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data; identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration; and classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

[0009] In accordance with an exemplary embodiment, a system is disclosed for recognizing gestures, comprising: a video camera for capturing at least one three-dimensional (3D) video stream of data on a subject; a module for extracting a time-series of skeletal data from the at least one 3D video stream of data; and a processor configured to: isolate a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data; identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration; and classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

[0010] In accordance with an exemplary embodiment, a non-transitory computer readable medium containing a computer program storing computer readable code is disclosed for recognizing gestures, the program being executable by a computer to cause the computer to perform a process comprising: capturing at least one three-

dimensional (3D) video stream of data on a subject; extracting a time-series of skeletal data from the at least one 3D video stream of data; isolating a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data; identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration; and classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

[0011] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory and are intended to provide further explanation of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The accompanying drawings are included to provide a further understanding of the invention, and are incorporated in and constitute a part of this specification. The drawings illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention.

[0013] FIG. 1 is an illustration of a gesture recognition system in accordance with an exemplary embodiment.

[0014] FIG. 2 is an illustration of a human skeleton system showing the body joints.

[0015] FIG. 3 is an illustration of a gesture recognition system in accordance with an exemplary embodiment.

[0016] FIG. 4 is an illustration of a flowchart illustrating a method of temporal segmentation for gesture analysis in accordance with an exemplary embodiment.

[0017] FIG. 5 is an illustration of a segmentation in accordance with an exemplary embodiment.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0018] It can be appreciated that when attempting to perform temporal segmentation of gestures, that is the task of finding gestures within a flow of human motion, numerous ambiguities can arise. For example, while some gestures can be performed subsequently without a pause in between (such gestures are called continuous gestures), some gestures include a pause right in the middle of the gesture, which can make it relatively impossible to trigger gesture boundaries by

simply observing abrupt changes from immobility to motion, or from motion to immobility.

[0019] Among change-points methods, Kernelized Temporal Cut (KTC) algorithm models the temporal segmentation problem as a series of two-sample problems within varying-size sliding windows, and solves it by using a test statistic based on Maximum Mean Discrepancy (MMD). In accordance with an exemplary embodiment, a method and system of temporal segmentation for gesture analysis is disclosed, which is referred herein as "Kinematic Kernelized Temporal Segmentation" (KKTS).

[0020] It can be appreciated that temporal segmentation, or the task of finding gestures within of a flow of human motion, can be crucial in many computer vision applications. For example, RGB-D sensors (or cameras), and their associated frameworks can provide for relatively easy and reliable extraction of a skeletal model from human users, and which can provide opportunities for the development of gesture recognition applications. However, temporal segmentation of gestures is still an open and challenging problem because it can be difficult to define a "gesture". Accordingly, it would be desirable to have a method and system for detecting gesture boundaries in an unsupervised and online manner while maintaining a decent tradeoff between over-segmentation and under-segmentation.

[0021] In accordance with an exemplary embodiment, a method and system of temporal segmentation for gesture analysis is disclosed, which is referred herein as a "Kinematic Kernelized Temporal Segmentation" (KKTS). For example, in accordance with an exemplary embodiment, while performing in real-time and in an unsupervised manner, the KKTS module (or algorithm) as disclosed herein, can locate gestures boundaries from a video-stream or flow of skeletal information. In addition, the KKTS module (or algorithm) can be used independent of any subsequent classification step or algorithm, which can make the system and method as disclosed herein, an ideal application for inclusion into a gesture processing system, including gesture recognition systems.

[0022] FIG. 1 is an illustration of a gesture recognition system 100 in accordance with an exemplary embodiment. As shown in FIG. 1, the system 100 can include a RGB-D camera 110 having, for example, Red, Green, Blue color space with a depth or distance capability, which can be used for acquiring color images (RGB color space) and a depth or distance of a subject or user 102 in each of the images. In

accordance with an exemplary embodiment, the subject or user 102 can be performing one or more gestures.

[0023] In accordance with an exemplary embodiment, the system 100 also preferably includes a segmentation and recognition system 120 and a display 130 having a graphical user interface (GUI) configured to display the results from the segmentation and recognition system 120. In accordance with an exemplary embodiment, the segmentation and recognition system 120 and/or the display 130 can include a computer or processing device having a memory, a processor, an operating system, one or more software applications for executing an algorithm as disclosed, and a display or graphical user interface (GUI) 130. It can be appreciated that the segmentation and recognition system 120 and/or the GUI or display 130 can be part of a standalone computer, or can be contained within one or more computer or processing devices.

[0024] FIG. 2 illustrates skeleton representation 200 for an exemplary user facing the RGB-D camera 120 wherein the skeleton 200 consists of 15 joints and 11 line segments representing head, shoulders and limbs of human body. As shown in FIG. 2, the line segments 210 are mutually connected by joints 220 and the movement of one segment is constrained by other. Furthermore, a few of the parts or line segments 210 can perform the independent motion while the others may stay relatively stationary, for example, such as a head movement.

[0025] In accordance with an exemplary embodiment, the position of a line segment 210 in 3D space can be determined by the two joints 220. For example, for a 3D skeleton frame, 15 body skeleton joints data can be extracted, which can be used to simulate the movement of human body.

[0026] FIG. 3 is an illustration of a gesture recognition system 300 in accordance with an exemplary embodiment. As shown in FIG. 3, the gesture recognition system 300 includes a data acquisition module 310, a temporal segmentation module 320, and a gesture recognition module 330.

[0027] In accordance with an exemplary embodiment, the data acquisition module 310 captures at least one three-dimensional (3D) video stream of data 312 on a subject performing one or more gestures. The 3D video stream of data can be obtained from, for example, a RGB-D camera 120, which is configured to capture RGB frames 312 and depth frames 314. In accordance with an exemplary embodiment, a time-series of skeletal data 318 is extracted from the at least one 3D

video stream of data based on a pose estimation 316 as disclosed herein. The time-series of skeletal data 318 can include, for example, a plurality of skeleton joints 220.

[0028] In accordance with an exemplary embodiment, the time-series of skeletal data 318 is input into the temporal segmentation module 320, which includes a KKTS module 322 having a KKTC module 324, which is configured to generate at least two temporal cuts 326. In accordance with an exemplary embodiment, the at least two temporal cuts 326 define non-overlapping adjacent segments partitioning the time-series of skeletal data 318. The temporal cuts 326 can then be input into the Cuts to Segment (CtS) module 328 of the KKTS module 322 to identifying segments containing gestures based on acceleration at each of the temporal cut 326. For example, if the rate of acceleration is positive at a temporal cut, the segment between the temporal cut and the consecutive temporal cut containing a gesture, for example, boundaries of a gesture boundary 340 can be recognized.

[0029] In accordance with an exemplary embodiment, the gesture recognition module 330 can receive the time-series of skeletal data 318 and the gesture boundaries 340, which can be input into a recognition algorithm or classification system 332 for determination of a recognized gesture 334.

[0030] FIG. 5 is an illustration of a flowchart 400 illustrating an exemplary method and system of temporal segmentation for gesture analysis, which can include a Kinematic Kernelized Temporal Cuts (KKTC) module 324, an optional hands-up decision function module 370, and Cuts to Segments (CtS) module 328.

[0031] In accordance with an exemplary embodiment, the input of skeleton joints 220 into the KKTS module 322 can be split into two method or algorithms, for example, a Kinematic Kernelized Temporal Cuts (KKTC) module 324 and a Cut to Segments (CtS) module 328. In accordance with an exemplary embodiment, the KKTC module 324 receives a time-series of skeletal data 318 of a user 102 performing gestures in front of the camera 120, and returns temporal cuts 326 as disclosed herein, which define non-overlapping adjacent segments partitioning the time-series of skeletal data 318. In accordance with an exemplary embodiment, the cuts to Segments (CtS) module 328 finds and returns, among all segments defined by temporal cuts, boundaries of segments containing gestures 340.

[0032] In accordance with an exemplary embodiment, a time-series of skeletal information or data 318 of size T, can be defined as $X \in R^{3N \times T}$. In accordance with

an exemplary embodiment, each element of *X* is a vector of *N* 3-dimensional skeleton joints 220, which are input into the KKTS module 324.

[0033] In accordance with an exemplary embodiment, the KKTS module 324 scans the sequences using two consecutive sliding windows of the same fixed size 350, 360. For example, the two consecutive sliding windows can be defined with $T_0 \in \mathbb{N}$, and $\delta T \in \mathbb{N}$, which can be two parameters respectively called *size of sliding windows* and *step length of moving the sliding windows*. For every t such that $t-T_0 \geq 1$ and $t+T_0-1 \leq T$, let $W_1^t = [\![t-T_0; t-1]\!]$ and $W_2^t = [\![t; t+T_0-1]\!]$, respectively the left and right sliding window at frame t.

[0034] In accordance with an exemplary embodiment, the two sliding windows can be used to compute an estimate of Maximum Mean Discrepancy (MMD) 350 within *X*. For example, the MMD 350 can be used to quantify global motion of body, and can be defined as follows:

$$MMD(t) = \frac{1}{T_0}\left[\sum_{i\in W_1^t}\sum_{j\in W_1^t} k(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i\in W_2^t}\sum_{j\in W_2^t} k(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_{i\in W_1^t}\sum_{j\in W_2^t} k(\mathbf{x}_i,\mathbf{x}_j)\right]$$

where *k* is a Gaussian kernel of bandwidth $\sigma = \sqrt{\frac{1}{2\gamma}}$, which can be defined as:

$$k(\mathbf{x},\mathbf{y}) = exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$$

which quantity or result, can be used in the KKTC module 324 to find rough location of temporal cuts 326.

[0035] In accordance with an exemplary embodiment, the KKTS module 322 can use the following kernelized kinematic quantities, defined with the same Gaussian kernel *k* as used in MMD:

- Global kernelized velocity of body at time *t*:

$$v(t) = 1 - \frac{1}{2T_v+1}\sum_{i=-T_v}^{T_v} k(\mathbf{x}_t,\mathbf{x}_{t+i})$$

although the calculated velocity is not used directly in the algorithm, it is used to describe the next two quantities. $T_v = 2$ can be a good value.

- Global kernelized acceleration of body at time *t*:

$$a(t) = v(t + T_a) - v(t - T_a)$$

Physically, it designates the rate of change of velocity with respect to time. $T_a = 1$ can be a good value. In accordance with an exemplary embodiment, it can be used by the CtS module 328 to find out which segments contain gestures.

- Global kernelized jerk of body at time *t*:

$$j(t) = v(t - T_j) - 2v(t) + v(t + T_j)$$

Physically, the global kernelized jerk of body designates the rate of change of acceleration with respect to time. $T_j = 4$ can be a good value. In accordance with an exemplary embodiment, the rate of change of acceleration (or the global kernelized jerk of body) can be used in the KKTC module 324 to find or locate a relatively precise location of the temporal cuts 326.

[0036] In accordance with an exemplary embodiment, an optional "hands-up" decision function (or module) 370 can also be used to assist with the identification of the temporal cuts 326 based on the assumption that a user is more likely to be in the middle of a gesture if the subject or users hands are up than they are down. For example, the following function, hereinafter called "hands-up" decision function denoted D can be defined as the sum of the vertical position of left-hand denoted L_y and vertical position of right-hand denoted R_y at time *t*, retrieved from *X*. The hands-up decision 370 can be expressed as follows:

$$D(t) = L_y(t) + R_y(t)$$

[0037] In accordance with an exemplary embodiment, the hands-up decision can be used in KKTC module 324 to refine location of the temporal cuts from rough to accurate.

[0038] In accordance with an exemplary embodiment, the quantities previously introduced to build both the KKTC module 324 and the CtS module 328 are further explained and once concatenated, the quantities can result in a finding of a gesture boundary 340.

[0039] In accordance with an exemplary embodiment, first, a local maxima of MMD along sliding windows providing rough location of temporal cuts is obtained. The amount of true positive and false negative cuts can be both decent, but the location of the cuts is approximate. Indeed, for example, the location of the cuts can tend to be too late at the beginning of a gesture and too early at the end. In parallel, local maxima of jerk estimate can be used to provide accurate location of cuts, but with false positives.

[0040] In accordance with an exemplary embodiment, each cut provided by a maximum of MMD can be refined to a cut provided by a local maximum of jerk , either forward or backward in time, using the value of the ”hands-up” decision function as disclosed herein. In accordance with an exemplary embodiment, this step assumes that a user is more likely to be in the middle of a gesture if its hands are up than if they are down.

[0041] In accordance with an exemplary embodiment, at the end of the process, the temporal cuts are both relevant and accurate, and few of them are false positives.

[0042] In accordance with an exemplary embodiment, the algorithm or steps performed by the KKTC module 324 are shown in Algorithm 1.

Algorithm 1 Kinematic Kernelized Temporal Cuts (KKTC) : isolate cuts delimiting segments potentially containing gestures out of a time-series of human motion

Input :

- $X = (\mathbf{x}_t) \in \mathbb{R}^{3N \times T}$ a sequence of N 3-dimensional skeleton joints of size T

Output :

- $\mathcal{C} \in [\![1; T]\!]^n$ a sequence of n cuts

Parameters:

- $T_0 \in \mathbb{N}$ the size of sliding windows
- $\delta T \in \mathbb{N}$ the step lenght of moving the sliding windows
- γ the parameter of the gaussian kernel

function $KS_{(T_0, \delta T, \gamma)}(X)$

Compute $\{\hat{t}_{MMD}\}$ the x-values of local maxima of $(MMD(t))_{t \in [\![1;T]\!] \setminus \delta T \mathbb{N}}$
Compute $\{\hat{t}_j\}$ the x-values of local maxima of $(j(t))_{t \in [\![1;T]\!]}$
$\mathcal{C} \leftarrow$ *empty list*

for $\hat{t} \in \{\hat{t}_{MMD}\}$ **do**

Find $\hat{t}_- \in \{\hat{t}_j\}$ such that $\hat{t}_- < \hat{t}$ and $|\hat{t}_- - \hat{t}|$ is minimal
Find $\hat{t}_+ \in \{\hat{t}_j\}$ such that $\hat{t}_+ > \hat{t}$ and $|\hat{t}_+ - \hat{t}|$ is minimal

if $D(\hat{t}_-) < D(\hat{t}_+)$ **then**
Find $\hat{t}_{corrected} \in \{\hat{t}_j\}$ such that $\hat{t}_{corrected} \leq \hat{t}_-$, $|\hat{t}_{corrected} - \hat{t}_-|$ is maximal, and $(D(t))_{t \in [\![\hat{t}_{corrected}, \hat{t}_-]\!]}$ is a sequence in strict ascending order

if $D(\hat{t}_-) > D(\hat{t}_+)$ **then**
Find $\hat{t}_{corrected} \in \{\hat{t}_j\}$ such that $\hat{t}_{corrected} \geq \hat{t}_+$, $|\hat{t}_{corrected} - \hat{t}_+|$ is maximal, and $(D(t))_{t \in [\![\hat{t}_+, \hat{t}_{corrected}]\!]}$ is a sequence in strict ascending order

if $D(\hat{t}_-) = D(\hat{t}_+)$ **then**
$\hat{t}_{corrected} \leftarrow \hat{t}$

$\mathcal{C}$.append($\hat{t}_{corrected}$)

return $\mathcal{C}$

[0043] Once the adjacent non-overlapping segments are identified by the KKTC module 324, the CtS module 328 is configured to identify segments containing gestures using acceleration. For example, in accordance with an exemplary embodiment, if the kernelized estimate of acceleration is positive at a cut position, then the segment between this cut and the next cut contains a gesture.

[0044] The algorithm or steps of the CtS module 328 are shown in Algorithm 2.

Algorithm 2 Cuts to Segments (CtS): transform a sequence of cuts delimiting segments potentially containing gestures to a sequence of segments containing gestures

Input :

- $X = (\mathbf{x}_t) \in \mathbb{R}^{3N \times T}$ a sequence of N 3-dimensional skeleton joints of size T
- $\mathcal{C} = (c_i) \in [\![1; T]\!]^n$ a sequence of n cuts

Output :

- $\mathcal{S} \subset [\![1; T]\!]^m$ a sequence of m non-overlapping segments

function $CtS(C)$

$\mathcal{C} \leftarrow [1, c_1, c_2, \ldots, c_{n-1}, c_n, T]$
$\mathcal{S} \leftarrow$ *empty list*

for each couple (c_i, c_{i+1}) of successive cuts in $\mathcal{C}$ **do**
  if $a(c_i > 0$ **then**
    $\mathcal{S}$.append$([\![c_i; c_{i+1}]\!])$

return $\mathcal{S}$

[0045] FIG. 5 is an illustration of a segmentation in accordance with an exemplary embodiment. As shown in FIG. 5, from top to bottom, synchronized RGB frames, skeleton frames, ground truth (manual) segmentation, and segmentation generated by KKTS are shown. Frames which belong to segments containing gestures have a crosshatched background. Frames which belong to segments not containing gestures have a white background. The presence of a gap between 2 represented frames means a cut occurred there. The figure represents two continuous gestures since there is no pause (i.e., no immobility phase) between them. In accordance with an exemplary embodiment, KKTS segments them correctly, and the generated segmentation does match the ground truth segmentation.

[0046] In accordance with an exemplary embodiment, a non-transitory computer readable medium containing a computer program storing computer readable code is disclosed for recognizing gestures, the program being executable by a computer to cause the computer to perform a process including: capturing at least one three-

dimensional (3D) video stream of data on a subject; extracting a time-series of skeletal data from the at least one 3D video stream of data; isolating a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data; identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration; and classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

[0047] The computer usable medium, of course, may be a magnetic recording medium, a magneto-optic recording medium, or any other recording medium which will be developed in future, all of which can be considered applicable to the present invention in all the same way. Duplicates of such medium including primary and secondary duplicate products and others are considered equivalent to the above medium without doubt. Furthermore, even if an embodiment of the present invention is a combination of software and hardware, it does not deviate from the concept of the invention at all. The present invention may be implemented such that its software part has been written onto a recording medium in advance and will be read as required in operation.

[0048] It will be apparent to those skilled in the art that various modifications and variation can be made to the structure of the present invention without departing from the scope or spirit of the invention. In view of the foregoing, it is intended that the present invention cover modifications and variations of this invention provided they fall within the scope of the following claims and their equivalents.

WHAT IS CLAIMED IS:

1. A method for recognizing gestures, comprising:

capturing at least one three-dimensional (3D) video stream of data on a subject;

extracting a time-series of skeletal data from the at least one 3D video stream of data;

isolating a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data;

identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration;

classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

2. The method of claim 1, comprising:

computing an estimated Maximum Mean Discrepancy (MMD) within the time-series of skeletal data; and

generating estimated temporal cuts among the time-series of skeletal data based on the estimated MMD.

3. The method of claim 2, comprising:

refining each of the estimated temporal cuts computed using the estimated MMD to generate a maximum rate of change of acceleration.

4. The method of claim 3, comprising:

generating the maximum rate of change of acceleration using a value of a hands-up decision function, wherein the hands-up decision function is a sum of vertical position of a left-hand joint and a right-hand joint at a time (t);

classifying a positive hands-up decision function a gesture; and

classifying a negative hands-up decision function as a non-gesture.

5. The method of claim 1, comprising:

classifying a positive rate of acceleration within a temporal cut as a beginning of the gesture; and

classifying a negative rate of acceleration within the temporal cut as an end of the gesture.

6 The method of claim 1, comprising:

inputting the time-series of skeletal data from the at least one 3D video stream of data and the gesture boundaries into a gesture recognition module; and

recognizing the gesture boundary as a type of gesture.

7. A system for recognizing gestures, comprising:

a video camera for capturing at least one three-dimensional (3D) video stream of data on a subject;

a module for extracting a time-series of skeletal data from the at least one 3D video stream of data; and

a processor configured to:

isolate a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data;

identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration;

classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

8. The system of claim 7, comprising:

a display for displaying results generated by the processor in which one or more gesture boundaries from the time-series of skeletal data in a visual format.

9. The system of claim 7, wherein the processor is configured to:

compute an estimated Maximum Mean Discrepancy (MMD) within the time-series of skeletal data; and

generate estimated temporal cuts among the time-series of skeletal data based on the estimated MMD.

10. The system of claim 9, wherein the processor is configured to:

refine each of the estimated temporal cuts computed using the estimated MMD to generate a maximum rate of change of acceleration.

generate the maximum rate of change of acceleration using a value of a hands-up decision function, wherein the hands-up decision function is a sum of vertical position of a left-hand joint and a right-hand joint at a time (t);

classifying a positive hands-up decision function a gesture; and

classifying a negative hands-up decision function as a non-gesture.

11. The system of claim 10, wherein the processor is configured to:

classify a positive rate of acceleration within a temporal cut as a beginning of the gesture; and

classify a negative rate of acceleration within the temporal cut as an end of the gesture.

12 The system of claim 7, comprising:

a gesture recognition module configured to receive the time-series of skeletal data from the at least one 3D video stream of data and the gesture boundaries, and recognizing the gesture boundary as a type of gesture.

13. The system of claim 7, wherein the video camera is a RGB-D camera, and wherein the RGB-D camera produces a time-series of RGB frames and depth frames.

14. The system of claim 7, wherein the module for extracting a time-series of skeletal data from the at least one 3D video stream of data and the processor are in a standalone computer.

15. A non-transitory computer readable medium containing a computer program storing computer readable code for recognizing gestures, the program being executable by a computer to cause the computer to perform a process comprising:

capturing at least one three-dimensional (3D) video stream of data on a subject;

extracting a time-series of skeletal data from the at least one 3D video stream of data;

isolating a plurality of points of abrupt content change and identifying each of the plurality of points of abrupt content change as a temporal cut, and wherein a plurality of temporal cuts define a set of non-overlapping adjacent segments partitioning the time-series of skeletal data;

identifying among the plurality of temporal cuts, temporal cuts of the time-series of skeletal data having a positive acceleration;

classifying each of the one or more pair of consecutive cuts with the positive acceleration as a gesture boundary.

16. The computer readable storage medium of claim 15, comprising:

computing an estimated Maximum Mean Discrepancy (MMD) within the time-series of skeletal data; and

generating estimated temporal cuts among the time-series of skeletal data based on the estimated MMD.

17. The computer readable storage medium of claim 16, comprising:

refining each of the estimated temporal cuts computed using the estimated MMD to generate a maximum rate of change of acceleration.

18. The computer readable storage medium of claim 15, comprising:

generating the maximum rate of change of acceleration using a value of a hands-up decision function, wherein the hands-up decision function is a sum of vertical position of a left-hand joint and a right-hand joint at a time (t);

classifying a positive hands-up decision function a gesture; and

classifying a negative hands-up decision function as a non-gesture.

19. The computer readable storage medium of claim 15, comprising:

classifying a positive rate of acceleration within a temporal cut as a beginning of the gesture; and

classifying a negative rate of acceleration within the temporal cut as an end of the gesture.

20 The computer readable storage medium of claim 15, comprising:

inputting the time-series of skeletal data from the at least one 3D video stream of data and the gesture boundaries into a gesture recognition module; and

recognizing the gesture boundary as a type of gesture.

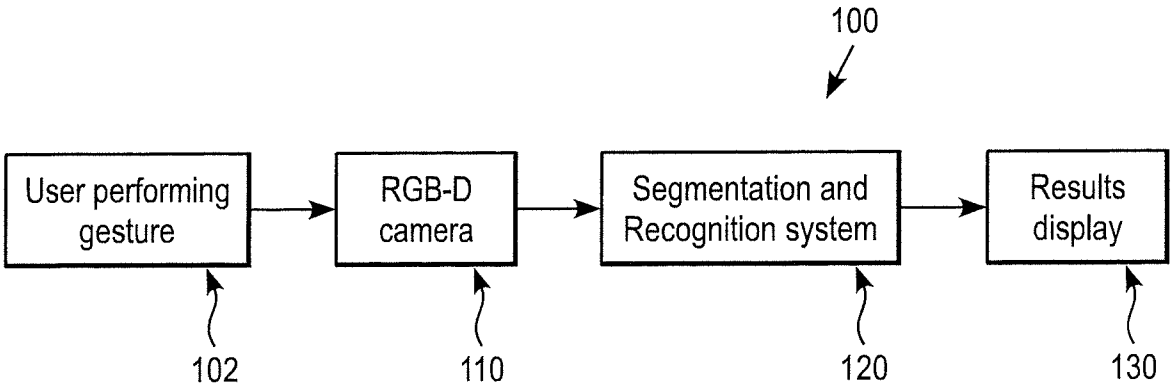

100
User performing gesture
RGB-D camera
Segmentation and Recognition system
Results display
102
110
120
130

FIG. 1

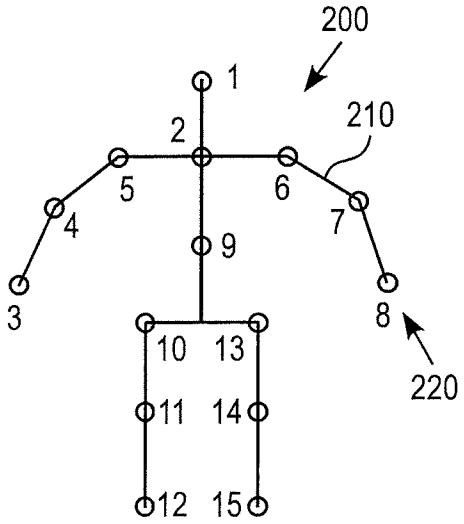

200
1
2
210
5
6
4
7
9
3
8
10
13
220
11
14
12
15

FIG. 2

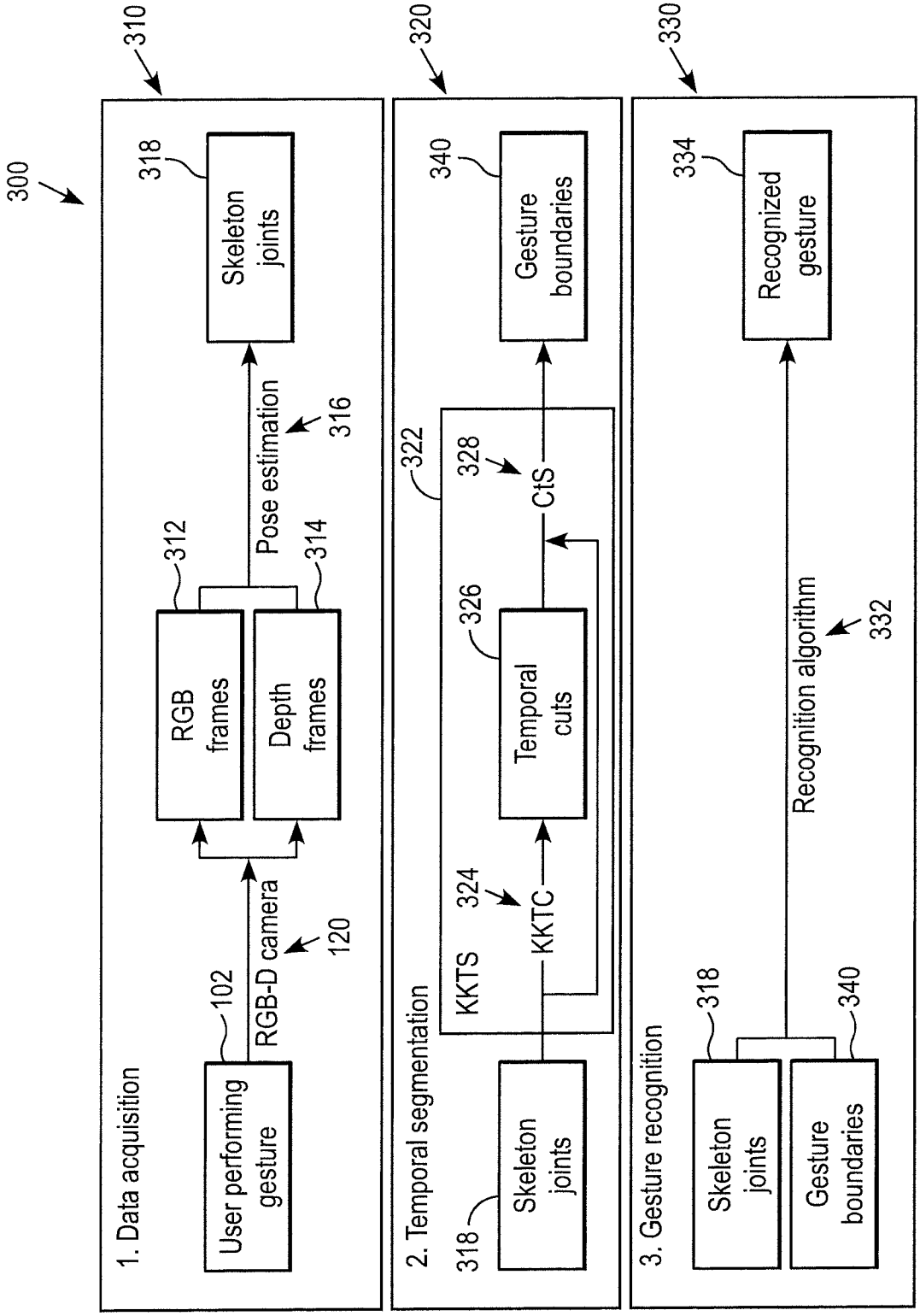

300
1. Data acquisition
310
User performing gesture
102
RGB-D camera
120
RGB frames
312
Depth frames
314
Pose estimation
316
Skeleton joints
318
2. Temporal segmentation
320
Skeleton joints
318
KKTC
324
KKTS
Temporal cuts
326
322
CtS
328
Gesture boundaries
340
3. Gesture recognition
330
Skeleton joints
318
Gesture boundaries
340
Recognition algorithm
332
Recognized gesture
334

FIG. 3

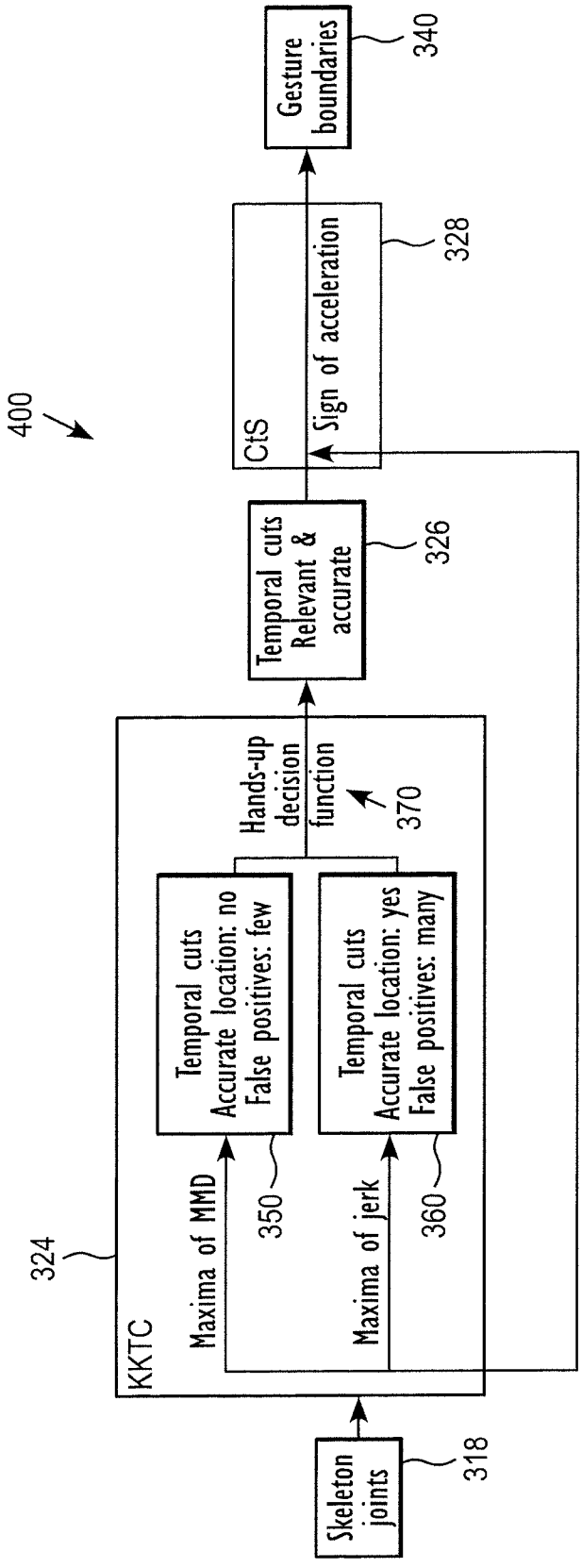
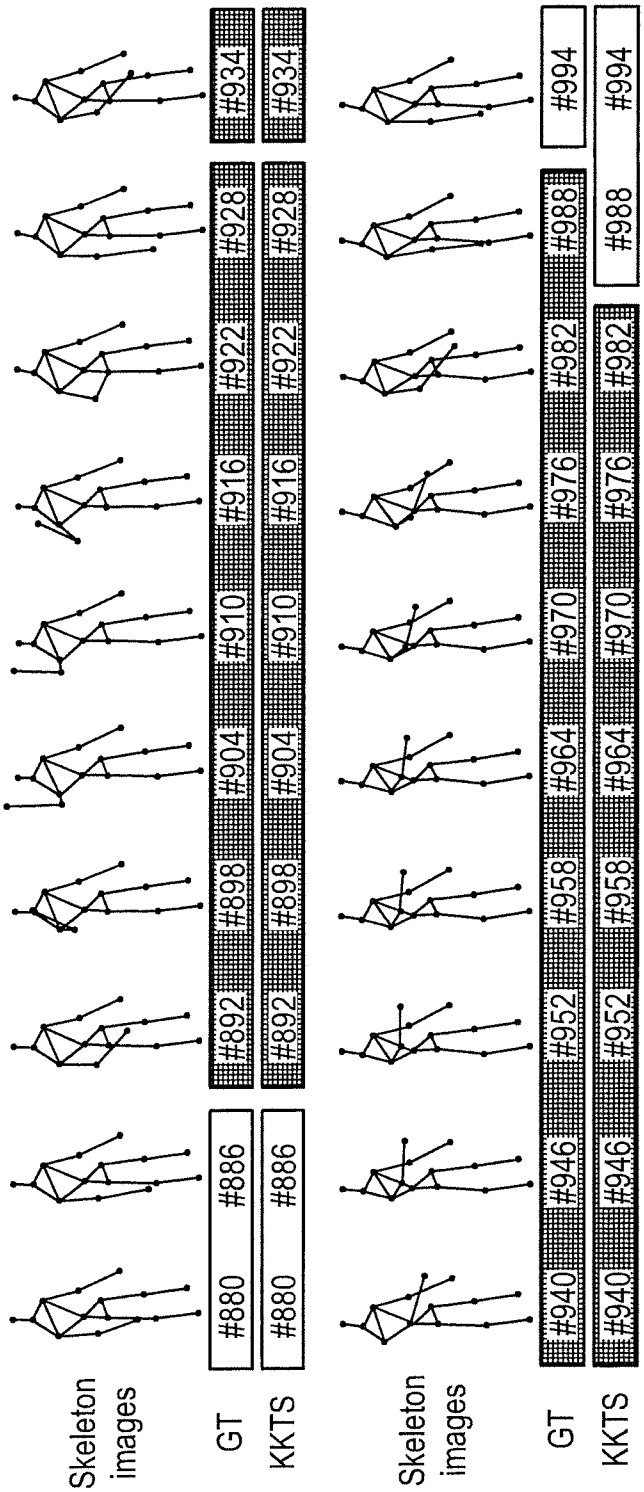

400
Skeleton joints
318
KKTC
324
Maxima of MMD
Maxima of jerk
Temporal cuts Accurate location: no False positives: few
350
Temporal cuts Accurate location: yes False positives: many
360
Hands-up decision function
370
Temporal cuts Relevant & accurate
326
CtS
Sign of acceleration
328
Gesture boundaries
340

FIG. 4

FIG. 5


Skeleton images
GT
KKTS
#880
#886
#892
#898
#904
#910
#916
#922
#928
#934
#940
#946
#952
#958
#964
#970
#976
#982
#988
#994

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US15/47095

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 3/01 (2015.01)
CPC - G06F 3/01
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)
IPC(8) Classification(s): G06F 3/01, 3/00; G06K 9/00 (2015.01)
CPC Classification(s): G06F 3/01, 3/017; G06K 9/00382, 9/00335, 9/00362

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)
PatSeer (US, EP, WO, JP, DE, GB, CN, FR, KR, ES, AU, IN, CA, Other Countries (INPADOC), RU, AT, CH, TH, BR, PH); IEEE/IEEEXplore; Google/Google Scholar; IP.com; Keywords: three dimensional, 3D, video, gesture, motion, movement, boundary, range, endpoint, start-point, acceleration, recognition, spotting, detection, modeling, sequence, series, time, interval, period, temporal, adjacent, skeletal, hand, finger, non-overlapping, separate, distinct, isolate, parse, extract, capture

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013/0034265 A1 (NAKASU, T et al.) February 07, 2013; abstract; paragraphs [0027], [0029], [0033], [0077]-[0079], [0104]	1, 5-8, 12, 14, 15, 19, 20
---		---
Y		2-4, 9-11, 13, 16-18
Y	US 2014/0067107 A1 (STANHOPE, S et al.) March 06, 2014; paragraphs [0039], [0053]	2-4, 9-11, 16, 17
Y	US 2011/0289455 A1 (REVILLE, B et al.) November 24, 2011; paragraphs [0004], [0119]-[0120], [0131]	4, 10, 11, 13, 18
A	US 2013/0294651 A1 (ZHOU, J et al.) November 07, 2013; entire document	1-20

☐ 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
"E" earlier application or patent but published on or after the international filing date
"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)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"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
"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
"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
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 October 2015 (26.10.2015)	27 NOV 2015
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)