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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 November 2016 (24.11.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/184705 A1

(51) International Patent Classification:
G06K 9/00 (2006.01) *G06T 7/00* (2006.01)
G06K 9/62 (2006.01) *A61B 5/024* (2006.01)

(21) International Application Number: PCT/EP2016/060250

(22) International Filing Date: 9 May 2016 (09.05.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
15168575.7 21 May 2015 (21.05.2015) EP

(71) Applicant: **KONINKLIJKE PHILIPS N.V.** [NL/NL]; High Tech Campus 5, 5656 AE Eindhoven (NL).

(71) Applicant *(for NL only)*: **TECHNISCHE UNIVERSITEIT EINDHOVEN** [NL/NL]; Den Dolech 2, 5612 AZ Eindhoven (NL).

(72) Inventors: **DE HAAN, Gerard**; High Tech Campus, Building 5, 5656 AE Eindhoven (NL). **WANG, Wenjin**; High Tech Campus, Building 5, 5656 AE Eindhoven (NL).

(74) Agents: **DE HAAN, Poul, Erik** et al.; High Tech Campus, Building 5, 5656 AE Eindhoven (NL).

(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))*

Published:

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

(54) Title: DETERMINING A PULSE SIGNAL FROM A VIDEO SEQUENCE

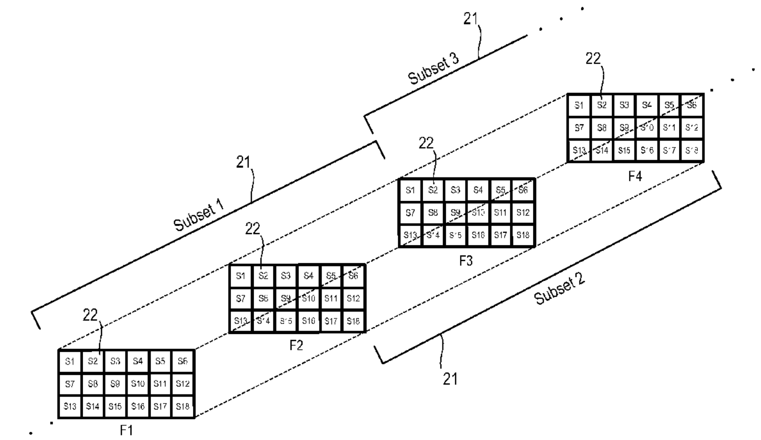

Figure 4(b)

(57) Abstract: According to an aspect, there is provided an apparatus for determining a pulse signal from a video sequence, the apparatus comprising a processing unit configured to obtain a video sequence, the video sequence comprising a plurality of image frames; form a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each image frame is divided into a plurality of frame segments; for a first video sub-sequence formed from frame segments from a first subset of image frames, comparing a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames; concatenate the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values; and determine a pulse signal from the concatenated video sub-sequences.

WO 2016/184705 A1

Determining a pulse signal from a video sequence

TECHNICAL FIELD OF THE INVENTION

The invention relates to an apparatus and method for determining a pulse signal from a video sequence.

BACKGROUND OF THE INVENTION

Recently, techniques for performing remote photoplethysmography (remote PPG or rPPG) have been developed. These techniques enable a PPG signal to be obtained from a video sequence of image frames captured using an imaging unit (e.g. a camera). It is desirable for a video sequence to be processed and rPPG signals extracted automatically so that subjects can be automatically monitored. However, this requires areas of living skin tissue to be automatically identified in the video sequence.

The task of detecting subjects in a video, as one of the fundamental topics in computer vision, has been extensively studied in the past decades. Given a video sequence containing subjects, the goal is to locate the regions corresponding to the body parts of a subject. Most existing work exploits human appearance features to discriminate between subject and background in a supervised training mechanism. However, a common problem with these methods is that their trained features are not unique to human beings, any feature that is similar to human appearance can be misclassified. Moreover, supervised methods are usually restricted to prior-known samples and tend to fail when unpredictable samples occur, e.g. a face detector trained with frontal faces cannot locate faces viewed from the side, while a skin classifier trained with bright skin subjects fails with dark skin subjects. Other types of methods require an area of skin to be manually selected in the video sequence, with this area being tracked over time to compensate for motion. However, this technique clearly requires manual input, and it is not easy to correctly track the selected area when there is substantial motion.

Based on the development of rPPG techniques, it has been observed that as compared to physical appearance features, the invisible physiological features (e.g. pulse) can better differentiate humans from non-humans in a video sequence. In the natural environment, only the skin tissue of an alive subject exhibits pulsatility, so any object that does not show a pulse-signal can be safely classified into the non-human category. This can prevent the false detection of objects with an appearance similar to humans, as shown, for example, in Figure 1.

Figure 1 provides two examples of how a living tissue detection technique should successfully operate. In the left hand image, a human face and an artificial face are present face on to the camera, and only the human face should be identified (despite the artificial face having similar physical appearance features to the human face), as indicated by the dashed box and outline of the area corresponding to living skin tissue. In the right hand image, a human face and an artificial face are present side on to the camera, and only the human face should be identified.

In the paper “Face detection method based on photoplethysmography” by G. Gibert and D. D’Alessandro, and F. Lance, 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 449-453, (2013) a hard threshold is set to select segmented local regions (e.g. grids, triangles or voxels) with higher frequency spectrum energy as skin-regions. In the paper “Automatic ROI detection for rPPG measurements” by R. van Luijtelaar, W. Wang, S. Stuijk, and G. de Haan, ACCV 2014, Singapore pre-defined clustering parameters are used to cluster regions sharing similarities as skin regions.

However, these methods still struggle to extract a useful pulse signal when there is significant movement of the subject or skin area.

Therefore it is an object to provide an improved method and apparatus for determining a pulse signal from a video sequence.

SUMMARY OF THE INVENTION

According to a first aspect, there is provided a method for determining a pulse signal from a video sequence, the method comprising obtaining a video sequence, the video sequence comprising a plurality of image frames; forming a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each subset comprises a plurality of image frames, wherein each image frame is divided into a plurality of frame segments, wherein each frame segment is a group of neighboring pixels in the image frame; for a first video sub-sequence formed from frame segments from a first subset of image frames, comparing a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames; concatenating the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values; and determining a pulse signal from the concatenated video sub-sequences.

In some embodiments, the method further comprising the steps of for the second video sub-sequence, comparing a representative value for the second video sub-sequence to representative values for video sub-sequences formed from frame segments from a third subset of image frames; and concatenating the first video sub-sequence and second video sub-sequence to a third video sub-sequence formed from frame segments from the third subset of image frames based on the comparison of representative values.

In some embodiments, the method further comprises repeating the steps of comparing and concatenating in order to concatenate video sub-sequences from each of a plurality of subsets of image frames.

In some embodiments, each subset of image frames comprises a respective set of consecutive image frames in the video sequence.

In some embodiments, each image frame is in more than one subset such that the subsets of image frames are overlapping.

In some embodiments, each subset comprises at least one image frame that is common to another subset.

In alternative embodiments, each image frame in the video sequence is in a respective subset of image frames.

In some embodiments, each subset of image frames comprises a respective set of three image frames in the video sequence.

In some embodiments, the first subset of image frames and the second subset of image frames comprise image frames that are adjacent in the video sequence.

In some embodiments, each frame segment is a group of neighboring pixels in the image frame.

In some embodiments, the frame segments have the same shape. In alternative embodiments, each image frame is divided into a plurality of frame segments by grouping pixels into frame segments based on color and spatial similarities of the pixels.

In some embodiments, each image frame is divided into a plurality of frame segments independently of the other image frames. In alternative embodiments, each image frame in a subset is divided into a plurality of frame segments based on the content of the image frame and the other image frames in the subset.

In some embodiments, the step of forming the plurality of video sub-sequences comprises forming each video sub-sequence from frame segments in corresponding spatial positions in the image frames in the subset of image frames.

In alternative embodiments, the step of forming the plurality of video sub-sequences comprises, for each video sub-sequence, selecting frame segments from each image frame in the subset of image frames such that a chromatic energy and/or spatial-distance energy between the frame segments in the video sub-sequence is minimized.

In some embodiments, the method further comprises the step of determining a representative value for each video sub-sequence.

In some embodiments, the step of determining a representative value for each video sub-sequence can comprise averaging the pixel values of pixels in the frame segments in the video sub-sequence.

In some embodiments, the step of averaging the pixel values can comprise weighting the pixel values of pixels in each frame segment, wherein the pixel values are weighted based on spatial position of the pixel in the frame segment and/or a difference in color with a pixel or group of pixels at or near the center of the frame segment; and averaging the weighted pixel values of pixels in a frame segment.

In some embodiments, the step of determining a representative value further comprises normalizing the average of the pixel values in the frame segment.

In some embodiments, the step of determining a representative value for each video sub-sequence comprises determining a difference between the averages of the pixel values of pixels in the frame segments in the video sub-sequence.

In some embodiments, the step of comparing comprises identifying the second video sub-sequence as a video sub-sequence formed from frame segments from the second subset of image frames that is similar or most similar to the first video sub-sequence.

In some embodiments, the step of comparing comprises identifying the second video sub-sequence as a video sub-sequence formed from frame segments from the second subset of image frames that is similar or most similar to the first video sub-sequence in spatial distance and/or representative value.

In some embodiments, the step of determining a pulse signal comprises determining the pulse signal from the representative values of the concatenated video sub-sequences.

In some embodiments, the steps of forming, determining a representative value, comparing, concatenating and determining a pulse signal are performed for image frames that are divided into a first plurality of frame segments; and wherein the method further comprises the steps of repeating the steps of forming, determining a representative value, comparing, concatenating and determining a pulse signal for the image frames in the video sequence when the image frames are divided into a second plurality of frame segments, the second plurality of frame segments comprising a different number of frame segments than the first plurality of frame segments.

In some embodiments, the method further comprises the steps of repeating the steps of comparing, concatenating and determining a pulse signal for other video sub-sequences formed from frame segments from the first subset of image frames to determine further pulse signals; and analyzing the pulse signals to identify areas of living skin tissue in the video sequence.

In some embodiments, the method further comprises the step of determining one or more physiological characteristics from one or more pulse signals associated with the identified areas of living skin tissue in the video sequence.

According to a second aspect, there is provided a computer program product comprising a computer readable medium having computer readable code embodied therein, the computer readable code being configured such that, on execution by a suitable computer or processor, the computer or processor is caused to perform any of the methods described above.

According to a third aspect, there is provided an apparatus for determining a pulse signal from a video sequence, the apparatus comprising a processing unit configured to obtain a video sequence, the video sequence comprising a plurality of image frames; form a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each subset comprises a plurality of image frames, wherein each image frame is divided into a plurality of frame segments, wherein each frame segment is a group of neighboring pixels in the image frame; for a first video sub-sequence formed from frame segments from a first subset of image frames, comparing a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames; concatenate the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values; and determine a pulse signal from the concatenated video sub-sequences.

In some embodiments, the processing unit is further configured to, for the second video sub-sequence, compare a representative value for the second video sub-sequence to representative values for video sub-sequences formed from frame segments from a third subset of image frames; and concatenate the first video sub-sequence and second video sub-sequence to a third video sub-sequence formed from frame segments from the third subset of image frames based on the comparison of representative values.

In some embodiments, the processing unit is further configured to repeat the comparison and concatenation in order to concatenate video sub-sequences from each of a plurality of subsets of image frames.

In some embodiments, each subset of image frames comprises a respective set of consecutive image frames in the video sequence.

In some embodiments, each image frame is in more than one subset such that the subsets of image frames are overlapping.

In some embodiments, each subset comprises at least one image frame that is common to another subset.

In alternative embodiments, each image frame in the video sequence is in a respective subset of image frames.

In some embodiments, each subset of image frames comprises a respective set of three image frames in the video sequence.

In some embodiments, the first subset of image frames and the second subset of image frames comprise image frames that are adjacent in the video sequence.

In some embodiments, each frame segment is a group of neighboring pixels in the image frame.

In some embodiments, the frame segments have the same shape. In alternative embodiments, each image frame is divided into a plurality of frame segments by grouping pixels into frame segments based on color and spatial similarities of the pixels.

In some embodiments, each image frame is divided into a plurality of frame segments independently of the other image frames. In alternative embodiments, each image frame in a subset is divided into a plurality of frame segments based on the content of the image frame and the other image frames in the subset.

In some embodiments, the processing unit is configured to form the plurality of video sub-sequences by forming each video sub-sequence from frame segments in corresponding spatial positions in the image frames in the subset of image frames.

In alternative embodiments, the processing unit is configured to form the plurality of video sub-sequences by, for each video sub-sequence, selecting frame segments from each image frame in the subset of image frames such that a chromatic energy and/or spatial-distance energy between the frame segments in the video sub-sequence is minimized.

In some embodiments, the processing unit is further configured to determine a representative value for each video sub-sequence.

In some embodiments, the processing unit is configured to determine a representative value for each video sub-sequence can by averaging the pixel values of pixels in the frame segments in the video sub-sequence.

In some embodiments, the processing unit is configured to average the pixel values by weighting the pixel values of pixels in each frame segment, wherein the pixel values are weighted based on spatial position of the pixel in the frame segment and/or a difference in color with a pixel or group of pixels at or near the center of the frame segment; and averaging the weighted pixel values of pixels in a frame segment.

In some embodiments, the processing unit is configured to determine a representative value by normalizing the average of the pixel values in the frame segment.

In some embodiments, the processing unit is configured to determine a representative value for each video sub-sequence by determining a difference between the averages of the pixel values of pixels in the frame segments in the video sub-sequence.

In some embodiments, the processing unit is configured to compare by identifying the second video sub-sequence as a video sub-sequence formed from frame segments from the second subset of image frames that is similar or most similar to the first video sub-sequence.

In some embodiments, the processing unit is configured to compare by identifying the second video sub-sequence as a video sub-sequence formed from frame segments from the second subset of image frames that is similar or most similar to the first video sub-sequence in spatial distance and/or representative value.

In some embodiments, the processing unit is configured to determine a pulse signal by determining the pulse signal from the representative values of the concatenated video sub-sequences.

In some embodiments, the processing unit is configured to form, determine a representative value, compare, concatenate and determine a pulse signal for image frames that are divided into a first plurality of frame segments; and the processing unit is further configured to repeat the forming, determining a representative value, comparing, concatenating and determining a pulse signal for the image frames in the video sequence when the image frames are divided into a second plurality of frame segments, the second plurality of frame segments comprising a different number of frame segments than the first plurality of frame segments.

In some embodiments, the processing unit is further configured to repeat the comparing, concatenating and determining a pulse signal for other video sub-sequences formed from frame segments from the first subset of image frames to determine further pulse signals; and analyze the pulse signals to identify areas of living skin tissue in the video sequence.

In some embodiments, the processing unit is further configured to determine one or more physiological characteristics from one or more pulse signals associated with the identified areas of living skin tissue in the video sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

For a better understanding of the invention, and to show more clearly how it may be carried into effect, reference will now be made, by way of example only, to the accompanying drawings, in which:

Figure 1 illustrates the desired operation of a living skin tissue detection technique;

Figure 2 is a block diagram of an apparatus according to an embodiment of the invention;

Figure 3 is a flow chart illustrating a method according to an embodiment of the invention;

Figures 4(a)-(d) illustrate how a pulse signal can be obtained from a video sequence;

Figure 4(a) illustrates how a video sequence is made up of a series of image frames;

Figure 4(b) illustrates how each of the image frames is divided into a plurality of frame segments;

Figure 4(c) illustrates how two video sub-sequences are formed using frame segments in the same spatial position within the image frames;

Figure 4(d) illustrates exemplary pulse signals for the two video subsequences so formed;

Figure 5 is a diagram illustrating the processing stages in the exemplary Voxel-Pulse-Spectral method;

Figure 6 illustrates segmentation of an image frame in three different scales;

Figure 7 illustrates four different measures of pairwise similarity and a resulting similarity matrix;

Figure 8 shows an example of similarity matrix decomposition using incremental sparse PCA; and

Figure 9 illustrates the projection of eigenvectors onto hierarchical voxels and a fused map indicating which parts of the video sequence correspond to living skin tissue.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

An apparatus 2 that can be used to determine a pulse signal from a video sequences according to an embodiment of the invention is shown in Figure 2. In further embodiments, the apparatus 2 can be used to identify living skin tissue from the determined pulse signal. The apparatus 2 comprises an imaging unit 4 that captures a video sequence over a period of time. The imaging unit 4 can be or comprise a camera, for example an RGB camera, that can be used for rPPG measurements. The imaging unit 4 provides a video sequence comprising a plurality of image frames to a processing unit 6.

The processing unit 6 controls the operation of the apparatus 2 and can comprise one or more processors, multi-core processors or processing modules for implementing the living skin tissue identification techniques described herein. In some embodiments, the processing unit 6 can be implemented as a plurality of processing modules, with each module being configured to perform a particular part or step of the living skin tissue identification techniques described herein.

The apparatus 2 further comprises a memory unit 8 for storing computer readable program code that can be executed by the processing unit 6 to perform the method according to the invention. The memory unit 8 can also be used to store or buffer the video sequence from the imaging unit 4 before, during and after processing by the processing unit 6 and any intermediate products of the processing.

It will be appreciated that in some embodiments the apparatus 2 can comprise a general-purpose computer (e.g. a desktop PC) with an integrated or separate imaging unit 4, or a portable computing device (e.g. a laptop, tablet or smart phone) that has an integrated or separate imaging unit 4. In some embodiments, the apparatus 2 can be dedicated to the purpose of determining a pulse signal from a video sequence, identifying living skin tissue in a video sequence using the pulse signal, and/or for measuring physiological characteristics of a subject from rPPG signals extracted from areas of a video sequence identified as corresponding to living skin tissue.

In practical implementations, the apparatus 2 may comprise other or further components to those shown in Figure 2 and described above, such as a user interface that allows a subject to activate and/or operate the apparatus 2, and a power supply, such as a battery or connection to a mains power supply, for powering the apparatus 2. The user interface may comprise one or more components that allow a subject to interact and control the apparatus 2. As an example, the one or more user interface components could comprise a switch, a button or other control means for activating and deactivating the apparatus 2 and/or pulse determination process. The user interface components can also or alternatively comprise a display, or other visual indicator (such as a light) for providing information to the subject about the operation of the apparatus 2. Likewise, the user interface components can comprise an audio source for providing audible feedback to the subject about the operation of the apparatus 2.

The flow chart in Figure 3 illustrates a method of determining a pulse signal from a video sequence according to an embodiment.

In step 101, an imaging unit 4 obtains a video sequence. The video sequence is made up of a series of image frames. A series of image frames 20 is shown in Figure 4(a).

Next, a plurality of video sub-sequences are formed from the image frames 20 (step 103). Each video sub-sequence comprises a frame segment from each of two or more image frames 20 in a subset of image frames 20 in the video sequence.

Each video sub-sequence is preferably formed from a frame segment from each of two or more consecutive image frames 20 (and thus each subset of image frames 20 preferably comprises consecutive image frames 20).

Figure 4(a) illustrates four consecutive image frames 20 from a video sequence, which are labelled F1, F2, F3, F4 respectively. Subsets 21 of image frames are formed that, in this example, each comprise three consecutive image frames 20. In addition, in this example, the subsets 21 are formed such that the subsets 21 are overlapping (i.e. it is possible for some or all of the image frames 20 to be part of more than one subset 21). Thus, Figure 4(a) shows a first subset 21, subset 1, that comprises image frames F1, F2 and F3, a second subset 21, subset 2, that comprises image frames F2, F3 and F4 and a third subset 21, subset 3, that comprises image frames F3, F4 and F5, etc. Each subset 21 therefore overlaps by two image frames 20 with the preceding adjacent subset 21.

It will be appreciated that in alternative embodiments, the subsets 21 can comprise a different number of image frames 20, and/or the extent of overlap between the subsets 21 can be different. For example, the subsets 21 can comprise three image frames 20, but each subset 21 may only overlap with the preceding adjacent subset 21 by one image frame 20. Alternatively, there may be no overlap between the image frames 20 in the subsets 21 (i.e. in this case each image frame 20 will only be part of a single subset 21).

Each image frame 20 is divided into a plurality of frame segments 22. Each frame segment 22 is a group of neighboring pixels in the image frame 20. An exemplary segmentation is illustrated in Figure 4(b) in which each frame 20 is divided into equal-sized squares. In alternative embodiments, the segments can be a different shape, such as triangular. In other (preferred) alternative embodiments, the shape of the segments 22 can be determined by the image in the image frame 20 (for example the boundaries of the shapes can follow boundaries between different colors in the image frame). In each embodiment, however, it will be appreciated that each frame segment 22 comprises a group of spatially-related (i.e. neighboring) pixels in each image frame 20. In the preferred embodiment, the frame segments 22 are also known in the art as ‘super pixels’, for example as described in “SLIC Superpixels Compared to State-of-the-art Superpixel Methods” by Achanta et al., IEEE Transactions on Pattern Analysis & Machine Intelligence 2012 vol.34 Issue No.11, Nov. 2012, pp: 2274-2282.

In Figure 4(b) each image frame 20 is divided into eighteen frame segments 22, and, for ease of reference, each frame segment 22 is given a respective label: S1, …, S18.

In the preferred ‘super pixel’ embodiment in which the grouping of pixels/shape of the segments 22 is determined by the content of the image frame, the frame segments 22 can be determined by grouping pixels in the image frame 20 based on color and spatial similarities of the pixels. In this way, neighboring or closely neighboring pixels having a similar or consistent color will be grouped together into a single frame segment 22.

In the above embodiments, an image frame 20 is divided into frame segments 22 based solely on an analysis of that image frame 20. However, it is possible in some embodiments to divide a particular image frame 20 into frame segments 22 based on an analysis of that image frame 20 and the other image frames 20 in that subset 21 of image frames. In other words, the spatial and/or color based segmentation of an image frame 20 described above is extended into the time domain so that pixels sharing appearance (e.g. color) and spatial similarities in the temporal domain are grouped together. This extension into the time domain can be used to result in the same frame segmentation pattern being applied to each image frame in the subset 21, or to each image frame in the subset 21 having a respective segment pattern.

As noted above, video sub-sequences are formed from a frame segment from each of the image frames 20 in a subset. These video sub-sequences are also referred to herein as ‘voxels’. Thus, referring to Figure 4(b), a plurality of video sub-sequences are formed from a frame segment 22 from each of the image frames 20 in subset 1, a plurality of video sub-sequences are formed from a frame segment 22 from each of the image frames 20 in subset 2, and so on.

In this illustrated embodiment, each frame segment 22 in each image frame 20 is part of a respective video sub-sequence 23, although in other embodiments this does not have to be the case (i.e. some frame segments 22 are not part of any video sub-sequences 23).

In some embodiments, which are illustrated in Figure 4(c), each video sub-sequence 23 is formed using frame segments 22 in the same spatial position within each image frame 20 in the subset 21. For example, one video sub-sequence 23 can be formed from the frame segment S1 in the top left corner of the image frames 20 in subset 1, another from the frame segment S13 in the bottom left corner of the image frames 20 in subset 1, and so on. Likewise, further video sub-sequences 23 are formed using frame segments 22 in the image frames 20 in subset 2. Figure 4(c) illustrates the sub-sequences 23 formed from the first and second subsets 21 according to this embodiment.

However, in a preferred embodiment (which is particularly preferred when the frame segments 22 are formed by grouping pixels according to color and spatial similarities), to improve the robustness of the method to changes in the content of the video sequence (for example due to a subject moving), a video sub-sequence 23 can be formed by selecting frame segments 22 from the image frames 20 in a subset 21 that are consistent with each other (for example generally consistent in spatial location within the image frames 20 and generally consistent in color). This can lead to the video sub-sequence ‘winding’ its way through the image frames 20 so that the video sub-sequence 23 contains frame segments 22 for a particular part of a subject (for example as a subject moves from left to right in the video sequence, a particular video sub-sequence 23 can be formed by a frame segment 22 in each image frame 20 that corresponds to the subject’s cheek due to spatial and color similarities). One preferred way to form the video sub-sequences 23 is, for a particular frame segment 22 in an image frame 20, to identify the frame segment 22 in the next image frame 20 that has the minimum chromatic energy (i.e. minimum difference in chrominance) and spatial-distance energy (i.e. minimum spatial-distance) from the particular frame segment 22. It will be understood that chromatic energy refers to an energy function based on the chrominance values of the pixels in the frame segment 22 and a frame segment 22 in the next image frame

20, and thus minimizing the chromatic energy to form a video sub-sequence 23 can comprise, for a particular frame segment 22, identifying the frame segment 22 in the next image frame 20 that has the smallest chromatic energy to the frame segment 22 under consideration. It will be appreciated that the more different the chrominance for the pixels in a frame segment 22 compared to the chrominance for the pixels in the frame segment 22 under consideration, the higher the chromatic energy, and thus the lower the likelihood that the frame segment 22 will be selected for that video sub-sequence 23. It will also be understood that spatial-distance energy refers to an energy function based on the spatial-position of the frame segment 22 in the image frame 20 and the spatial-position of the frame segment 22 in the next image frame 20, and thus minimizing the spatial-distance energy to form a video sub-sequence 23 can comprise, for a particular frame segment 22, identifying the frame segment 22 in the next image frame 20 that provides the smallest spatial distance in the frame segment 22 under consideration. It will be appreciated that the larger the distance from the position of a frame segment 22 in the next image frame 20 to the position of the frame segment 22 under consideration, the higher the spatial-distance energy, and the lower the likelihood that the frame segment 22 will be selected for that video sub-sequence 23. In some embodiments, the spatial-distance energy can be considered after projecting RGB pixel values to a subspace (e.g. UV). In an alternative approach, it is also possible to initialize the new voxels at time t using the centers of the voxels 23 at time t-1. In cases where the voxels 23 overlap in time (i.e. overlapping image frames), the center in the last frame segment 22 of a voxel 23 determines the center in the first frame segment 22 of the next voxel 23.

The division of the image frames 20 into the plurality of frame segments 22 described above provides video sub-sequences at a first scale or resolution, with the scale or resolution being indicated by the number of frame segments 22 in each image frame (e.g. eighteen in Figure 4), or by the size of each frame segment 22 (e.g. in terms of the number of image pixels per frame segment 22). In some embodiments the image frames 20 can also be divided into a second plurality of frame segments 22 having a different size/resolution to the first plurality, with a further plurality of video sub-sequences 23 being formed from those segments 22. This use of multiple resolutions has the advantage that it enables scale-invariant detection of a subject in a video sequence and it improves the detection of multiple subjects in a video sequence.

Once the video sub-sequences 23 are formed, the method proceeds to identify video sub-sequences 23 that can be concatenated or linked together in order to allow a pulse signal to be determined. The method makes use of the fact that a pulse signal in living skin tissue of a particular subject is more or less identical at different places of the skin (provided that normalized color values are considered), which means that video sub-sequences 23 from (slightly) different spatial parts of the video sequence can be concatenated together.

Thus, in step 105, a first video sub-sequence 23 formed from frame segments 22 from a first subset 21 of image frames 20 is compared to video sub-sequences 23 formed from frame segments from a second subset 21 of image frames 20. The second subset 21 is preferably the subset adjacent the first subset 21.

The comparison preferably comprises comparing at least one representative value for the first video sub-sequence 23 to respective representative values for video sub-sequences 23 in the second subset 21. A representative value for a video sub-sequence is based on or derived from the content (e.g. pixel values) of the frame segments 22 that make up the video sub-sequence 23. In some embodiments, each video sub-sequence 23 can have more than one representative value. In this case, step 105 can comprise comparing multiple ones of the representative values for a particular video sub-sequence 23 to multiple ones of the representative values for other video sub-sequences 23.

Figure 4(d) illustrates an example of step 105. Thus, a first video sub-sequence 24 formed from frame segments 22 in image frames 20 in a first subset 21 (subset 1 comprising image frames F1, F2 and F3), and in particular a sub-sequence 24 formed from segment S9 in image frames F1, F2 and F3, is compared to video sub-sequences 25 formed from frame segments 22 in image frames 20 in a second subset 21 (subset 2 comprising image frames F2, F3 and F4). In some embodiments, the first video sub-sequence 24 can be compared to all of the video sub-sequences 25 formed from the second subset 21 of image frames 20. In alternative embodiments, the first video sub-sequence 24 can be compared to only some of the video sub-sequences 25 formed from the second subset 21 of image frames 20. For example the first video sub-sequence 24 may only be compared to the video sub-sequences 25 that are in the same and/or similar spatial positions in the image frames 20. This is illustrated in Figure 4(d) by the first video sub-sequence 24 only being compared with the labelled video sub-sequences 25 in frames F2, F3 and F4, i.e. the video sub-sequences 25 comprising segments S2, S3, S4, S8, S9, S10, S14, S15 and S16.

The comparison in step 105 aims to identify a video sub-sequence 25 that the first video sub-sequence 24 can be concatenated with, and as such, the comparison aims to identify a or the video sub-sequence 25 that has a representative value that is similar or most similar to the representative value of the first video sub-sequence 24. In some embodiments, the video sub-sequence 25 should also be in a similar spatial position in the image frames 20 to the first video sub-sequence 24 (e.g. as illustrated in Figure 4(d)).

It will be appreciated that when comparing representative values to decide upon which sub-sequences to concatenate, embodiments in which there is an overlap in time of the sub-sequences (e.g. as with subset 1 and subset 2 in Figure 4) provide a benefit, i.e. it is possible to compare the representative values of each sub-sequence in the same frame (in the case of averages) or derived from the same frame-pair (in the case of a (normalized) difference).

Prior to the step of comparing, the representative values for each video sub-sequence 23 can be determined. In Figure 4, the representative value for a particular video sub-sequence 23 formed from frame segment SX in image frames Y to Z in a video sequence is denoted $V_{SX:Y \rightarrow Z}$. The annotation used for the representative value V for each video sub-sequence 23 in Figure 4(c) is shown beneath the video sub-sequences themselves.

In some embodiments, a representative value for each video sub-sequence 23 can be determined from an average of the pixel values (e.g. RGB values) of pixels in all of the frame segments 22 in the video sub-sequence 23. In alternative embodiments, where each video sub-sequence 23 may have more than one representative value, each value can be the average of the pixel values of pixels in a respective frame segment 22 in the video sub-sequence 23.

In some embodiments, when determining the average, the pixel values can be weighted based on spatial position of the pixel in the frame segment 22 and/or a difference in color with a pixel or group of pixels at or near the center of the frame segment 22, and the average of the weighted values determined. For example the pixel values can be weighted based on the distance from the pixel to the spatial boundary, and/or the distance from the pixel to the center of the frame segment 22. Preferably the weighting leads to the pixels close to the segment 22 boundary being less weighted as they are less reliable than pixels close to the middle of the segment 22 due to jittering artefacts between neighboring segments 22.

In some embodiments the average of the pixel values determined above is normalized, and the normalized average is used as the representative value. The average can be normalized by dividing the average by an average (or weighted average, if desired) of the pixel values of all the pixels in all of the frame segments 22 in the video sub-sequence 23. Alternatively, the average can be normalized by dividing the average by an average (or weighted average, if desired) of the pixel values of all of the pixels in all of the image frames 20 in the subset 21 from which the video sub-sequence 23 is formed.

In some (or further) embodiments, a representative value for each video sub-sequence 23 can be determined from a difference between the averages (or weighted averages) of the pixel values of pixels for each frame segment in the video sub-sequence 23. In some embodiments the differences can be normalized using the average (or weighted average) of pixel values across all frame segments 22 in the sub-sequence 23. Alternatively the differences can be normalized using a sum of the average (or weighted average) of pixel values across all frame segments 22 in the sub-sequence 23. The use of normalized differences is advantageous as it allows the difference values (i.e. video sub-sequences) to be concatenated and a pulse signal computed even when the average values representing different video sub-sequences are significantly different. Where normalized differences are concatenated, the concatenated difference values can be integrated before or after a pulse signal is derived (step 109 below).

Returning to Figure 3, in step 107, based on the comparison between the first video sub-sequence 24 and the video sub-sequences 25 from the second subset 21, the first video sub-sequence 24 is concatenated with a video sub-sequence 25 from the second subset 21 that is sufficiently similar (e.g. the representative values are within a particular amount of each other) or most similar to the first video sub-sequence 24 (this video sub-sequence 25 is referred to herein as the second video sub-sequence 25).

This concatenation is shown in Figure 4(e), where the first video sub-sequence 24 (formed from segment S9 in F1, F2, F3) is concatenated with second video sub-sequence 25 (formed from segment S10 in F2, F3, F4).

Once the first video sub-sequence 24 has been concatenated with a second video sub-sequence 25 that is formed from frame segments from the second subset 21 of image frames 20, steps 105 and 107 can be repeated for the second video sub-sequence 25 to identify a third video sub-sequence 26 that is formed from frame segments from the third subset 21 of image frames 20 (i.e. F3, F4, F5). Thus, the third video sub-sequence 26 can be concatenated with the second video sub-sequence 25 as shown in Figure 4(e). Steps 105 and 107 can then be repeated to identify video sub-sequences from further subsets 21 of image frames 20 that are to be concatenated with the sub-sequences 24, 25, 26.

Once the video sub-sequences 23 have been concatenated or linked together, a pulse signal is determined from the video sub-sequences (step 109). The pulse signal represents the color, or changes in the color, of the frame segments 22 and/or video sub-sequences 23 in the concatenated sequence.

In some embodiments step 109 can comprise determining the pulse signal from the representative values of each of the concatenated video sub-sequences 24, 25, 26. Thus, in some embodiments, the pulse signal can be formed from the representative values themselves, for example as shown in Figure 4(f). Thus, the representative values $V_{S9:1\rightarrow3}$, $V_{S10:2\rightarrow4}$ and $V_{S...:3\rightarrow5}$ provide the values of the pulse signal 27 at times t_{F1}, t_{F2}, t_{F3} respectively (corresponding to the timing of the first image frame represented in each video sub-sequence 24, 25, 26).

In this example, the video sub-sequences 24, 25, 26 forming the pulse signal 27 contain an area of living skin tissue, and therefore the pulse signal 27 determined from this concatenated set of video sub-sequences will exhibit characteristics typical of a PPG signal (i.e. varying in amplitude consistent with changes in blood perfusion in the skin of the subject due to the beating of the heart). If the concatenated set of video sub-sequences did not contain an area of living skin tissue, the pulse signal determined from this set will not exhibit characteristics that are typical of a PPG signal (and, in the absence of changes in the ambient lighting in the video sequence, the pulse signal for this set may correspond generally to a noise signal).

It will be appreciated that in embodiments where the pulse signal 27 is formed from the representative values of the concatenated video sub-sequences 23, steps 107 and 109 can effectively be viewed as one step, as the or a representative value for the first video sub-sequence 24 is concatenated with the or a representative value for the second video sub-sequence 25.

In alternative embodiments, the concatenated video sub-sequences 24, 25, can be analyzed using techniques for extracting a pulse signal from a sequence of images/image frames/frame segments. Various techniques for determining a pulse signal from a video sequence are known in the art and will not be described in detail herein. However, some exemplary techniques are mentioned in the description of the Voxel-Pulse-Spectral method presented below.

In some embodiments, after forming a concatenated set of video sub-sequences and determining a pulse signal 27, the method can return to step 105 and repeat for a different one of the video sub-sequences 23 that is formed from the image frames 20 in the first subset, subset 1. In this way, a plurality of pulse signals 27 will be determined for different (spatial) parts of the video sequence, some of which may include areas of living skin tissue, whereas others may not. It will be appreciated that due to the way in which video sub-sequences 23 from different subsets 21 are compared to each other (i.e. by a comparison of representative values), it is possible that a particular video sub-sequence 23 can contribute to several different pulse signals 27 (or indeed a particular video sub-sequence 23 may not contribute to any pulse signals 27).

Once a pulse signal 27 (or a plurality of pulse signals) has been determined or extracted, the pulse signal 27 can be analyzed to identify areas of living skin tissue in the video sequence. Where video sub-sequences 23 are formed from frame segments having multiple resolutions, the analyzing can comprise analyzing the pulse signals obtained in different resolutions together to identify the areas of living skin tissue. In some cases, the pulse signals 27 can be clustered together based on similarities (e.g. spatial, temporal, color and/or frequency similarities), and the areas of living skin tissue identified from those clusters.

In some embodiments, the pulse signal 27 or pulse signals 27 can be analyzed to determine which, if any, exhibit characteristics of living skin tissue. In some embodiments frequency characteristics of the pulse signal(s) is/are determined, and the determined frequency characteristics of the pulse signal(s) is/are compared to typical frequency characteristics of pulse signals obtained from areas of living skin tissue. For example, a fixed frequency threshold or band (e.g. corresponding to a typical heart beat/pulse frequency) can be used to determine whether a periodic pulse signal corresponds to living skin tissue.

In alternative embodiments, the pulse signals 27 are spatially clustered based on similarities in the pulse signals. A suitable clustering algorithm for implementing this is density-based spatial clustering of applications with noise (DBSCAN), or the clustering proposed in “Face detection method based on photoplethysmography” that is referenced above.

However, in some embodiments, pairwise similarities for each pulse signal 27 with the other pulse signals 27 can be determined. That is, for each pulse signal, a measure of the similarity with each of the other pulse signals is determined. These pairwise similarity measures are then analyzed to identify areas of living skin tissue.

The pairwise similarity measures preferably include or are frequency-based pairwise similarity measures. This is advantageous since different frame segments 22 corresponding to areas of living skin tissue of the same particular subject should exhibit pulse signals that have similar (or the same) frequency peak index, phase or a low entropy in the correlation.

The measure of pairwise similarity for a pulse signal and one of the other pulse signals can be a measure of the correlation between at least part of the frequency spectra of the pulse signal and the one of the other pulse signals (which is referred to herein as the spectrum peak), a measure of the normalized cross-correlation between at least part of the frequency spectra of the pulse signal and the one of the other pulse signals (which is referred to herein as the spectrum phase), a measure of the regularity of correlation between at least part of the frequency spectra of the two pulse signals (which is referred to herein as the spectrum entropy) and/or the result of an inner product of the two pulse signals (which can optionally be filtered before the inner product is calculated). Further details of these pairwise similarity measures can be found in the description of the Voxel-Pulse-Spectral (VPS) method below.

Those skilled in the art will be aware of alternative or further measures of pairwise similarity to those presented above that can be determined and used to identify areas of living skin tissue.

In some embodiments, multiple measures of pairwise similarity (e.g. spectrum peak and spectrum phase) can be determined for each pulse signal with each of the other pulse signals, and those measures combined to form a distance metric representing the pairwise similarity for each pulse signal with each of the other pulse signals.

In some embodiments, a similarity matrix for the video sequence can be determined, and areas of living skin tissue can be identified from the similarity matrix. The similarity matrix can be formed by combining the pairwise similarities (or distance metrics) determined for the pulse signals. The similarity matrix is a matrix in which similar pulse signals across the two or more pluralities of video sub-sequences are mutually correlated. The use of a similarity matrix is advantageous as it does not require any parameters to be predefined (e.g. parameters based on skin tone or clustering).

Once the similarity matrix is determined, areas of living skin tissue are identified by performing matrix decomposition of the similarity matrix. In some embodiments, the matrix decomposition can by singular value decomposition (SVD), QR decomposition, sparse SVD, incremental SVD, principal component analysis, PCA, or independent component analysis, ICA. These techniques are generally known in the art and will not be described in detail herein. In a preferred embodiment, which is described in more detail below, the similarity matrix is decomposed using incremental sparse PCA.

The decomposition can comprise factorizing (decomposing) the similarity matrix into orthogonal bases to find the parts of the video sequence belonging to the same subject. This factorizing results in different kinds of similarities being separated into independent directions. This results in the frame segments belonging to the same subject being clustered in the same direction.

Once areas of living skin tissue have been identified in the video sequence, one or more physiological characteristics of the subject (the subject with the identified living skin tissue) can be determined from the video sequence.

In some embodiments, the physiological characteristic(s) can be determined from the one or more pulse signals 27 associated with the identified areas of living skin tissue. In this case, the one or more pulse signals 27 can be individually analyzed to determine a physiological characteristic and the physiological characteristics combined (e.g. averaged) to give an overall measure of the physiological characteristic for the subject. Alternatively the one or more pulse signals 27 can be combined (e.g. averaged) to give a single pulse signal, and the pulse signal analyzed to determine a physiological characteristic.

In other embodiments, the video sequence can be re-processed to extract a pulse signal or signals from the areas identified to be living skin tissue, and that pulse signal(s) processed to determine the physiological characteristic.

The pulse signal(s) derived from the video sequence are similar to signals obtained using a PPG sensor, so the one or more physiological characteristics of the subject can include any characteristic that can be derived from a PPG signal or other measure of the blood perfusion (or changes in the blood perfusion) of the subject, such as heart rate, heart rate variability, beat-to-beat interval, breathing rate, breathing signal, SpO2 value (i.e. the arterial oxygenation level of the blood), etc. Thus, techniques known in the art for deriving such physiological characteristics from PPG signals (e.g. peak detection in the frequency domain for determining heart rate) can be used to derive values for physiological characteristics from the pulse signal(s) obtained from areas identified to be living skin tissue.

A particular embodiment of the techniques presented herein is described below, and is referred to herein as the Voxel-Pulse-Spectral (VPS) method.

<u>Voxel-Pulse-Spectral (VPS) method</u>

<u>Camera-based pulse extraction</u>

In the human cardiovascular system, blood pulses propagating throughout the body changes the blood volume in skin tissue. Since the optical absorption of haemoglobin in blood varies across the light spectrum, detecting color variations of skin reflection can reveal the pulse rate. Recent remote photoplethysmography (rPPG) techniques enable the detection of pulse-induced color variations on human skin using a regular RGB camera. Blind Source Separation methods (e.g. PCA-based and ICA-based) have been proposed to factorize the temporal RGB signals for finding the pulse. A Chrominance-based rPPG method has also been proposed to define the pulse as a linear combination of RGB channels under a standardized skin-tone assumption, which is one of the most accurate rPPG methods in dealing with realistic challenges (e.g. different subject skin colors).

Pulse-based region of interest detection

Given the fact that the human pulse can be measured by rPPG in video sequences, the pulse signal can thus be used to assist the subject detection, i.e. detecting alive subjects by locating their living skin tissue. An existing technique proposes a face detection method based on the pulse signal which slices the video into fixed rigid-grids for local pulse extraction. It sets a hard threshold to find the grids with a high spectrum energy and label them as the face region. It is limited to videos in which the stationary face needs to be placed at a pre-defined distance from the camera. The VPS method described herein does not suffer from these limitations. In another existing technique, an optimal Region of Interest (RoI) selection method on the face to enhance the rPPG monitoring was proposed. However, the RoI is constrained to predefined facial landmarks, which is not a general solution for subject detection, i.e. it cannot detect other body parts (e.g. hands) that might be visible in a video. In contrast, the VPS method described herein does not make such an assumption and can detect all body parts with pulsatile blood volume.

An overview of the VPS method is shown in Figure 5, which takes an input video sequence comprising image frames 20 and outputs the subject RoI 52 (i.e. the areas of the video sequence that correspond to living skin tissue). Consistent with the method shown in Figure 3, the video sequence is segmented into a plurality of sub-sequences (hierarchical voxels stage 54 – step 103), pulse signals are determined from the sub-sequences (pulse extraction stage 56 – steps 105-109), a similarity matrix is determined and analyzed to identify RoIs (spectral analysis stage 58). Each stage is discussed in detail below.

Hierarchical voxels

Given a video sequence without any prior information about the subject (or about the content in general), the video sequence is first segmented into dense local regions where a pulse can be independently measured (this is the hierarchical voxels stage 54 in Figure 5). Although the video sequence can be sliced into fixed rigid-grids, this means that the subject size is quantized by the grid geometry, which struggles or fails when the subject is small or when there is body motion. Therefore, in the VPS method, it is preferred to use a superior video segmentation method for pulse extraction which is called 'hierarchical voxels'. The hierarchical voxels consist of spatiotemporally coherent clusters (frame segments), in multiple scales (with the scale determining the number of clusters/segments that each image frame 20 is divided into), where pixels in image frames 20 sharing appearance and spatial similarities in the temporal domain are grouped together. Although multiple scales are preferred and used in the following description of the VPS method, it is possible to use a single scale in the VPS method.

Starting from one scale, constructing the voxels (video sub-sequences that comprise a frame segment from a subset (e.g. 2, 3, 4, etc.) of consecutive image frames) is defined as the procedure of minimizing the chromatic energy E_c (i.e. minmizing the difference in chrominance for the frame segments) and spatial-distance energy E_s (i.e. minimizing the difference in spatial distance between the frame segments) between temporally-adjacent superpixels/frame segments in a short interval $T \in \{2n+1, n \in \mathbb{N}^+\}$ as:

$$\arg\min\Big(\sum_{t-\frac{T-1}{2}}^{t+\frac{T-1}{2}}\sum_{p\in P(t)}(1-\lambda)E_c^t(p,k)+\lambda E_s^t(p,k)\Big), \quad (1)$$

where $p \in P(t)$ is the set of pixels in the t-th frame. The representation of p is a 4-dimensional feature vector (x, y, u, v), where (x, y) and (u, v) are respectively the coordinates in the image plane and the chromatic plane (e.g. UV plane of YUV space, the empirical space for skin segmentation). K-means clustering is performed to assign pixels into k clusters for minimizing the total energy during T. λ is the parameter controlling the balance between two energy terms.

Furthermore, the single scale voxels are extended to multiple scales by initializing different k in equation 1 simultaneously, where each scale is an independent clustering. Considering that the voxels in separate scales have different resolutions and energy variations, the λ_i in i-th scale is adaptively self-tuned based on its own energy variations at t as:

$$\lambda_i^t = \log(k)\sqrt{\frac{\sigma(\phi(u_i^t))\cdot\sigma(\phi(v_i^t))}{\sigma(\phi(x_i^t))\cdot\sigma(\phi(y_i^t))}}, \quad (2)$$

where σ(·) denotes the standard deviation operator; $\phi(\cdot)$ represents the set of cluster means; *log(k)* controls the voxel compactness, i.e. voxels with higher resolution (larger k) should be more compact. The real-time tuning of λ in different scales avoids volatile and flickering clustering, which preserves the fine-grained segmentation.

There are four benefits of using hierarchical voxels as described above, as illustrated in Figure 6, which shows three different resolutions/scales (k = 16, k = 36; k = 64) for two different images comprising a live subject and an artificial face. In Figure 6, the first image 60 corresponds to the left hand image in Figure 1 and shows a human face and an artificial face that are both facing the camera. The second image 62 corresponds to the right hand image in Figure 1 and shows a human face and artificial face side on to the camera. Firstly, using hierarchical voxels establishes the spatio-temporally coherent “tubes” (video sub-sequences) for pulse measurement. Secondly, it enables the scale-invariant subject detection in a video. Thirdly, it maintains high boundary recall of subject shapes (as indicated by the boundaries of the frame segments following shapes in the image. Fourthly, it creates a statistical observation of skin-regions, since the pulse measured from voxels with different resolutions have different quantized qualities.

Pulse extraction

This section describes pulse extraction stage 56 in Figure 5. Each voxel (i.e. video sub-sequence) in the hierarchy is assumed to be an independent pulse sensor in parallel pulse extraction. In a preferred embodiment, the Chrominance-based method (CHROM) is used for pulse measurement, which is described in “Robust pulse rate from chrominance-based rPPG” by G. de Haan and V. Jeanne, TBME, 60(1):2878-2886, 2013. However, those skilled in the art will be aware of other techniques that can be used to determine the pulse signal for a voxel. For example, as noted above, techniques based on PCA or ICA can be used to extract a pulse signal from a voxel, or the “PBV” method as described in “Improved motion robustness of remote-PPG by using the blood volume pulse signature” by G. de Haan and A. van Leest, Physiol. Meas. 35 1913, 2014, can be used to extract a pulse from RGB traces. Different from CHROM that uses the spatially averaged RGB of all pixels to derive the pulse signal in a local region/cluster, the pixel RGB values in a voxel are combined by weighting them based on their distance to the voxel boundary, i.e. pixels close to the voxel boundary are less reliable due to occasional jittering artefacts between neighboring voxels and thus should be less weighted. Assuming that the closest Euclidean distance from a pixel k to the voxel boundary is d_k, the average RGB of j-th voxel in i-th scale at t is combined as:

$$(\bar{R}_{ij}^t, \bar{G}_{ij}^t, \bar{B}_{ij}^t) = \frac{\sum_{k=0}^{N} \left(d_k \cdot (R_{ijk}^t, G_{ijk}^t, B_{ijk}^t)\right)}{\sum_{k=0}^{N} d_k}, \quad (3)$$

where N denotes the number of pixels in j-th voxel. In a constant lighting environment, human skin tissue shows the same relative PPG-amplitude, but the chromatic differences in voxels lead to the variations in pulse-amplitudes. So different from CHROM, the temporal derivatives of average RGB in a voxel are used, i.e. $dC_{ij}^t = C_{ij}^t - C_{ij}^{t-1}, C \in \{\bar{R}, \bar{G}, \bar{B}\}$ to derive its chrominance-signals. In the interval T (defined in equation 1), the normalized chrominance derivatives are calculated as:

$$\begin{cases} \overrightarrow{dX}_{ij}^T = 3\frac{\overrightarrow{dR}_{ij}^T}{\sum_{t=0}^{T} dR_{ij}^t} - 2\frac{\overrightarrow{dG}_{ij}^T}{\sum_{t=0}^{T} dG_{ij}^t} \\ \overrightarrow{dY}_{ij}^T = 1.5\frac{\overrightarrow{dR}_{ij}^T}{\sum_{t=0}^{T} dR_{ij}^t} + \frac{\overrightarrow{dG}_{ij}^T}{\sum_{t=0}^{T} dG_{ij}^t} - 1.5\frac{\overrightarrow{dB}_{ij}^T}{\sum_{t=0}^{T} dB_{ij}^t} \end{cases}, \quad (4)$$

where $(dR_{ij}^t, dG_{ij}^t, dB_{ij}^t)$ denotes the temporal derivatives of RGB in a voxel between two image frames. The chrominance derivatives estimated in each interval are linearly combined into pulse derivatives and further integrated. Afterwards, different pulse intervals are overlap added to a complete pulse-signal $\overrightarrow{S}_{ij}^L$ with length L. This procedure is interpreted as:

$$\overrightarrow{S}_{ij}^L = \sum_{t=0}^{L-T+1} \overrightarrow{S}_{ij}^{t+T} + w \cdot \mathbf{csum}(\overrightarrow{dX}_{ij}^T - \frac{\sigma(\overrightarrow{dX}_{ij}^T)}{\sigma(\overrightarrow{dY}_{ij}^T)} \overrightarrow{dY}_{ij}^T), (5)$$

where csum($\cdot$) denotes the cumulative sum of temporal derivative signals; w is the Hanning window for smoothing the overlap adding. Consequently, the parallel extracted pulse-signals $\overrightarrow{S}_{ij}^L$ (from j-th voxel in i-th scale) are centralized and normalized as:

$$\overrightarrow{S}_{ij}^L = \frac{\overrightarrow{S}_{ij}^L - \mu(\overrightarrow{S}_{ij}^L)}{\sigma(\overrightarrow{S}_{ij}^L)}, \quad (6)$$

where $\mu(\cdot)$ denotes the averaging operation. Note that the pulse-signal is the only feature used in this method. No other appearance features like color or texture are used.

Spectral analysis

This section describes spectral analysis stage 58, which comprises three sub-stages, forming a similarity matrix, performing incremental sparse PCA on the similarity matrix, and using hierarchical fusion to identify areas of living skin tissue in the video sequence.

It has been noted that pulse signals extracted from skin regions belonging to the same subject share similarities in many aspects such as phase and frequency, whereas the ones extracted from non-skin regions (e.g. background) are random noises without correlation. Therefore, after extracting the pulse signals from hierarchical voxels, pairwise similarities of pulse signals are used to find alive subjects. This is also applicable to the case of multiple subjects in the video sequence, because the pulse measured from different subjects can be differentiated in phase and frequency as well.

Similarity matrix – In this step, a similarity matrix $\Sigma = (D,C)$ is created to interconnect the hierarchical voxels based on the measured pulse. In Σ, the entries D in the diagonal trace contain all voxels in different scales; the remaining entries C denote the pairwise connection between any pair of voxels. To build such a similarity matrix, the distance metric for measuring the pulse similarity needs to be defined. The most commonly used distance metrics, i.e. L1 and L2 distances, are not applicable to the pulse feature. However, compared to other appearance features (e.g., Haar and HOG), an essential and unique character in the pulse feature is that it contains periodicity. It has been noted that pulse signals from the same subject show the following relations: (1) they have similar frequency and thus their cross-correlation presents a significant spectrum peak; (2) they have no significant phase shift; (3) their frequency correlation is regular and less disordered; and (4) if considering pulse signals as multidimensional vectors, the included angle between two similar vectors is small. Therefore, a preferred distance metric used to build the similarity matrix for pulse signals emphasizes the above connections, and is composed of four different measurements:

Spectrum peak – In the frequency domain, a pulse-rate band $f \in [40, 240]$ BPM (Beats Per Minute) is defined for voxels to communicate, which is a broad range for healthy subjects including neonates and sporting subjects. The spectrum peak of two cross-correlated pulse-signals is defined as:

$$F = \arg \max_{f \in [40,240]} \left(\mathcal{F}(\overrightarrow{S}_{ij}^{L}) \circ \mathcal{F}(\overrightarrow{S}_{i'j'}^{L})^{*} \right), \qquad (7)$$

where $\circ$ denotes the element-wise product; * is the conjugation; $\mathcal{F}(\cdot)$ represents the Fast Fourier Transform (FFT).

Spectrum phase – Two similar pulse signals are also in the same phase, so their normalized cross-correlation should show a strong response in the time domain as:

$$P = \max\left(\mathcal{F}^{-1}(NCC)\right), \tag{8}$$

with

$$NCC = \frac{\mathcal{F}(\overrightarrow{S}^{L}_{ij}) \circ \mathcal{F}(\overrightarrow{S}^{L}_{i'j'})^{*}}{\left\|\mathcal{F}(\overrightarrow{S}^{L}_{ij}) \circ \mathcal{F}(\overrightarrow{S}^{L}_{i'j'})^{*}\right\|_2}, \tag{9}$$

where $\|\cdot\|_2$ is the L2-norm; $\mathcal{F}^{-1}(\cdot)$ denotes the inverse FFT.

Spectrum entropy – The term "entropy" is used to measure the regularity of correlation between two pulse-signals as:

$$E = \frac{\sum_{f=40}^{240} NCC(f) \log\left(NCC(f)\right)}{\log(240-40)}, \tag{10}$$

where the interpretation of E is consistent with the other measurements, i.e. larger E denotes better correlation.

Inner product – In the time domain, we use the inner product to measure the cosine angle between two pulse-signals as:

$$I = < \frac{\overrightarrow{S}^{L}_{ij}}{\|\overrightarrow{S}^{L}_{ij}\|_2}, \frac{\overrightarrow{S}^{L}_{i'j'}}{\|\overrightarrow{S}^{L}_{i'j'}\|_2} >, \tag{11}$$

where <,> denotes the inner product operation.

Finally, these four measurements are normalized to the range [0; 1] and fused together with a Gaussian kernel as:

$$\Sigma = 1 - \exp\left(-\frac{(F \circ P \circ E \circ I)^2}{2\sigma^2_{I,F,P,E}}\right), \tag{12}$$

where $\sigma_{I,F,P,E}$ represents the entry-wise standard deviation between four matrices. It should be noted that the four measurements are not completely independent from each other, the redundancy between measurements is beneficial for reducing the uncertainty in similarity estimation. Figure 7 shows an example of four measurements and their fused similarity matrix Σ 64, 66 for two video sequences, one that includes a single living subject, and the other including two living subjects. The entries with higher energy represent the index of similar voxels in the hierarchy.

In the distance metric used herein, two well-aligned pulse signals show boosted frequency energy during the cross correlation, which can effectively suppress the noise entries (e.g. voxels without pulse). In contrast, previous distance metrics are all objective measurements that cannot enhance the connection between similar entries in the comparison. In the end, all voxels in the hierarchy are mutually connected in the similarity matrix. The task of detecting an alive subject in voxels can be reformulated as finding a subspace partition of the similarity matrix such that the entries in the same subspace have identical similarity direction.

Incremental sparse matrix decomposition – The similarity matrix Σ 64, 66 can be interpreted as a linear combination of $\lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T + \ldots \lambda_n x_n x_n^T$, where $x_i \in X$ is a set of orthogonal vectors in the multi-dimensional space. In order to find the voxels belonging to the same subject, a matrix decomposition technique is used to factorize Σ into *X*, where different subjects are separated into different eigenvectors. Since Σ is a sparse matrix with many zero entries (e.g. the voxels pointing at background share no similarity), sparse PCA is applied to decompose Σ into *X* by seeking a trade-off between expressive power and data interpretability.

Sparse PCA is described in "Sparse PCA: Convex relaxations, algorithms and applications" by Y. Zhang, A. d'Aspremont and L. Ghaoui, International Series in Operations Research & Management Science Volume 166:915-940, 2012. The Sparse PCA finds the first sparse eigenvector with the maximum variance in Σ by optimizing the following non-convex objective function:

$$\arg\max_X (X^T \Sigma X) \text{ subj.to } \|X\|_2 = 1, \|X\|_1 \leq n, \quad (13)$$

where $\|\cdot\|_2$ is the L1-norm; $n > 0$ controls the cardinality of *X*. However, computing sparse eigenvectors with maximum variance is a combinatorial problem and numerically hard to solve, so the non-convex rank constraint in equation 13 is dropped following the lifting procedure for semidefinite relaxation with l_1 penalization as:

$$\arg\max_{\hat{\Sigma}} \mathbf{Tr}(\Sigma\hat{\Sigma}) - \rho\|\hat{\Sigma}\|_1 \text{ subj.to } \mathbf{Tr}(\hat{\Sigma}) = 1, \hat{\Sigma} \succeq 0, \quad (14)$$

where Tr(·) denotes the matrix trace operation; $\rho > 0$ controls the sparsity; $\hat{\Sigma} = XX^T$ is a symmetric matrix approximated by the first leading eigenvector. At this point, an algorithm named Hybrid Conditional Gradient Smoothing (HCGS) can be used to solve equation 14. HCGS is described in "Hybrid conditional gradient-smoothing algorithms with applications to sparse and low rank regularization" by A. Argyriou, M. Signoretto and J. Suykens, arXiv preprint: 1404.3591, 2014. The merit of HCGS is the fast convergence in convex relaxation using conditional gradient approaches.

However in practice, Σ may consist of multiple sparse eigenbasis in case of multiple subjects, whereas equation 14 only promotes the sparsity in the first leading eigenvector. To address this issue, the succeeding sparse eigenvectors x_i are estimated by sequentially deflating Σ with preceding sparse eigenvectors $x_1, x_2, ..., x_{i-1}$ using Hotelling's deflation as:

$$\Sigma_i = \Sigma_{i-1} - (x_i^T \Sigma_{i-1} x_i) x_i x_i^T,\ i \in [1, m], \qquad (15)$$

with

$$m = \arg\max_i \left(\frac{x_{i-1}^T \Sigma_{i-1} x_{i-1}}{1 + x_i^T \Sigma_i x_i} \right), \qquad (16)$$

where $x_i \in X$ can be derived by the power iteration in HGCS; m is the automatically found number of most expressive eigenvectors, which also implies the number of subjects in a video sequence, i.e. m is usually found at the largest eigenvalue gap.

Figure 8 shows an example of similarity matrix decomposition using incremental sparse PCA for the two similarity matrices 64, 66 shown in Figure 7, where similar voxels are factorized into the same directions in the selected eigenvectors. Thus Figure 8 shows the factorized and selected eigenbasis from a similarity matrix; the noisy entries in the original Σ 64, 66 are eliminated in $\hat{\Sigma}$ 68, 70; the eigenvalues (shown in graphs 72 and 74 respectively) clearly show the number of most expressive eigenvectors in $\hat{\Sigma}$ (shown in graphs 76 and 78 respectively).

As a matter of fact, some intrinsic (e.g. pulse rate variation) and extrinsic (e.g. luminance changes) factors may occasionally change the similarity matrix in subsequent frames, which leads to an instability of the sparse eigenvectors estimated from each single frame. To solve this problem, incremental subspace updating is employed to smoothly adapt the $x_i \in X$ to real-time changes in the time domain. Basically, it considers the time-varying similarity matrix $\hat{\Sigma}_{new}$ as a new observation, and use multiple observations $[\hat{\Sigma}_{old}, \hat{\Sigma}_{new}]$ from different frames to enrich the subspace model as:

$$[U, D, V] = \mathbf{SVD}([\hat{\Sigma}_{old}, \hat{\Sigma}_{new}]), \qquad (17)$$

where SVD(·) denotes the Singular Value Decomposition; U and D are incrementally updated eigenvectors and eigenvalues. An exemplary algorithm to incrementally estimate multiple sparse eigenvectors from a time-varying similarity matrix is shown in Algorithm 1 below:

Input: similarity matrix $\Sigma \in \mathbb{R}^{n \times n}$, eigenvectors U, eigenvalues D

$\rho = 0.25$ (sparsity), $N = 100$ (iteration times)
for $k = 1, 2, ..., N$ **do**
  $\beta_k = \frac{1}{n}\mathbf{Tr}(\Sigma\Sigma_k) + \frac{\rho}{n}\|\Sigma_k\|_1$
  $Z_k = \Sigma - \frac{\sqrt{k}\rho}{2\sqrt{2}}\Sigma_k + \frac{\sqrt{k}\rho}{2\sqrt{2}}\text{sign}(\Sigma_k) \odot (|\Sigma_k| - \frac{2\sqrt{2}}{n\sqrt{k}})$
  $X = \{x_1, x_2, ..., x_m\} \leftarrow$ multiple eigenvectors of Z_k, where m is determined by Eq. 15 and Eq. 16
  $\hat{\Sigma} = XX^T$
  $\Sigma_{k+1} = (1 - \frac{2}{k+1})\Sigma_k + \frac{2}{k+1}\hat{\Sigma}$
  if $\frac{|\beta_k - \beta_{k-1}|}{\beta_k} < 10^{-3}$ **then**
    break
  end if
end for
if $U, D == 0$ **then**
  $[U, D, V] = \mathbf{SVD}(\hat{\Sigma})$
else
  $[U, \hat{\Sigma}'] R = \mathbf{QR}([UD, \hat{\Sigma}]) \leftarrow$ solved by QR decomposition
  $[U', D', V'] = \mathbf{SVD}(R)$
  $U'_m \in U'$, $D'_m \in D'$ $\leftarrow$ select top m eigenvectors and eigenvalues, where m is determined by Eq. 16
  $U = \text{sign}(U) \odot |[U', \Sigma'_{new}]U'_m|$, $D = D'_m \leftarrow$ update subspace model
end if

Output: updated U and D

Algorithm 1

Hierarchical fusion – By projecting the estimated sparse eigenvectors 76, 78 onto the hierarchical voxels, a voxel-based human objectness map in multiple scales is obtained, where each scale has a different quantized description to the subject. This projection for the video sequence containing two subjects from Figures 7 and 8 is illustrated in Figure 9. The eigenvector 78 not only decides the subject direction (sign) in subspaces, but also decides the pulsatility (amplitude) of corresponding skin-regions, i.e., forehead and cheek show relatively high pulsatility in projection. The final step is to fuse multiple objectness maps into a single output. Due to the fact that hierarchical measurement creates a statistical observation for skin-regions, the basic idea is to exploit this redundancy to derive a single output for which all scales have the highest agreement. In this sense, the hierarchical fusion is cast as the energy minimization between multiscale objectness maps as:

$$\arg\min_{\hat{o}}(\gamma E_1 + (1-\gamma)E_2), \qquad (18)$$

with

$$\begin{cases} E_1 = \sum_i \sum_j \|o_{ij}, \hat{o}\|_2 \\ E_2 = \sum_i (\sum_{\substack{j \\ o_{ij} \subseteq o_{i-1,j}}} \|o_{ij}, o_{i-1,j}\|_2 + \sum_{\substack{j \\ o_{i+1,j} \subseteq o_{ij}}} \|o_{ij}, o_{i+1,j}\|_2) \end{cases}, \qquad (19)$$

where o_{ij} corresponds to the objectness value of *j*-th voxel in *i*-th scale that determined by the eigenvector elements; $\hat{o}$ denotes the fused objectness map; γ controls the balance between two energy terms. In equation 19, E_1 minimizes the energy between inputs and output, while E_2 minimizes the energy between spatially overlapped voxels in different scales, i.e., an implicit tree structure. Figure 9 shows an example of the fused result in the video sequence with two alive subjects, where separate outputs 80, 82 are provided for each of the identified subjects.

The above method, and the preferred embodiment of the VPS method, provide an improved method and apparatus for identifying living skin tissue in a video sequence. In particular the method provides improved living skin tissue detection rates compared to conventional techniques, with the detection being based solely on using pulse signals to detect living tissue. These improvements are obtained regardless of the scale (i.e. distance from the imaging unit 4), posture, position, skin tone, visible body part or motion of the subject, or background of the subject in the video sequence, whether the subject is partially occluded from the imaging unit 4, whether there is an artificial face or body part present in the video sequence, or whether there are multiple living subjects in the video sequence.

Variations to the disclosed embodiments can be understood and effected by those skilled in the art in practicing the claimed invention, from a study of the drawings, the disclosure and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality. A single processor or other unit may fulfil the functions of several items recited in the claims. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage. A computer program may be stored/distributed on a suitable medium, such as an optical storage medium or a solid-state medium supplied together with or as part of other hardware, but may also be distributed in other forms, such as via the Internet or other wired or wireless telecommunication systems. Any reference signs in the claims should not be construed as limiting the scope.

CLAIMS:

1. An apparatus for determining a pulse signal from a video sequence, the apparatus comprising:

a processing unit configured to:

obtain a video sequence, the video sequence comprising a plurality of image frames;

form a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each subset comprises a plurality of image frames, wherein each image frame is divided into a plurality of frame segments, wherein each frame segment is a group of neighboring pixels in the image frame;

for a first video sub-sequence formed from frame segments from a first subset of image frames, comparing a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames;

concatenate the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values; and

determine a pulse signal from the concatenated video sub-sequences.

2. An apparatus as claimed in claim 1, the processing unit being further configured to repeat the comparison and concatenation in order to concatenate video sub-sequences from each of a plurality of subsets of image frames.

3. An apparatus as claimed in claim 1 or 2, wherein each image frame is divided into a plurality of frame segments by grouping pixels in the image frame into frame segments based on color and spatial similarities of the pixels.

4. An apparatus as claimed in any of claims 1-3, wherein the processing unit is configured to form the plurality of video sub-sequences by, for each video sub-sequence, selecting frame segments from each image frame in the subset of image frames such that a difference between the chrominance of the frame segments and/or spatial-distance between the position of the frame segments in the video sub-sequence is minimized.

5. An apparatus as claimed in any of claims 1-4, wherein the processing unit is further configured to determine a representative value for each video sub-sequence by averaging the pixel values of pixels in the frame segments in the video sub-sequence.

6. An apparatus as claimed in claim 5, wherein the processing unit is configured to average the pixel values by:

weighting the pixel values of pixels in each frame segment, wherein the pixel values are weighted based on spatial position of the pixel in the frame segment and/or a difference in color with a pixel or group of pixels at or near the center of the frame segment; and

averaging the weighted pixel values of pixels in a frame segment.

7. An apparatus as claimed in claim 5 or 6, wherein the processing unit is further configured to determine a representative value by normalizing the average of the pixel values in the frame segment.

8. An apparatus as claimed in any of claims 5-7, wherein the processing unit is configured to determine a representative value for each video sub-sequence by determining a difference between the averages of the pixel values of pixels in the frame segments in the video sub-sequence.

9. A method for determining a pulse signal from a video sequence, the method comprising:

obtaining a video sequence, the video sequence comprising a plurality of image frames;

forming a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each subset comprises a plurality of image frames, wherein each image frame is divided into a plurality of frame segments, wherein each frame segment is a group of neighboring pixels in the image frame;

for a first video sub-sequence formed from frame segments from a first subset of image frames, comparing a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames;

concatenating the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values; and

determining a pulse signal from the concatenated video sub-sequences.

10. A method as claimed in claim 9, the method further comprising repeating the steps of comparing and concatenating in order to concatenate video sub-sequences from each of a plurality of subsets of image frames.

11. A method as claimed in claim 10, wherein each image frame is divided into a plurality of frame segments by grouping pixels into frame segments based on color and spatial similarities of the pixels.

12. A method as claimed in claim 10 or 11, wherein the step of forming the plurality of video sub-sequences comprises, for each video sub-sequence, selecting frame segments from each image frame in the subset of image frames such that a difference between the chrominance of the frame segments and/or spatial-distance between the position of the frame segments in the video sub-sequence is minimized.

13. A method as claimed in claim 12, the method further comprising the step of determining a representative value for each video sub-sequence by averaging the pixel values of pixels in the frame segments in the video sub-sequence.

14. A method as claimed in claim 13, wherein the step of averaging the pixel values comprises:

weighting the pixel values of pixels in each frame segment, wherein the pixel values are weighted based on spatial position of the pixel in the frame segment and/or a difference in color with a pixel or group of pixels at or near the center of the frame segment; and

averaging the weighted pixel values of pixels in a frame segment.

15. A computer program product comprising a computer readable medium having computer readable code embodied therein, the computer readable code being configured such that, on execution by a suitable computer or processor, the computer or processor is caused to perform the method of any of claims 9-14.

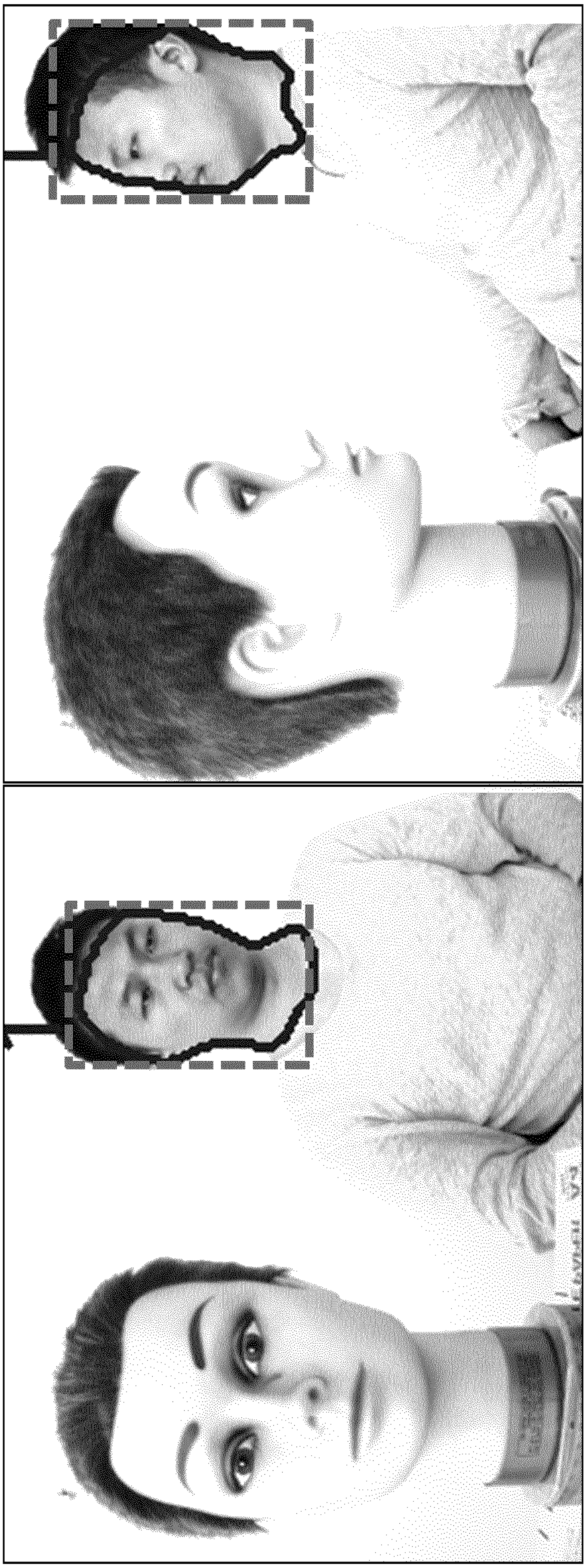

Figure 1

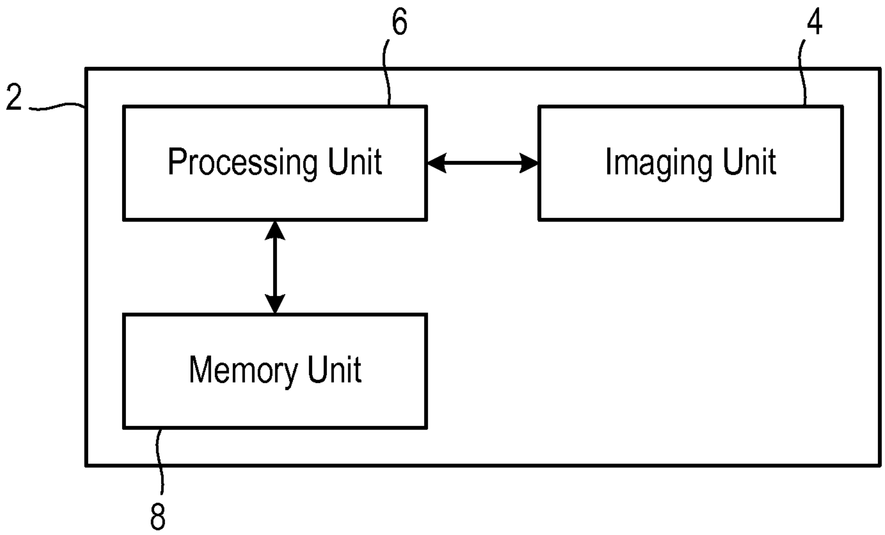

Figure 2

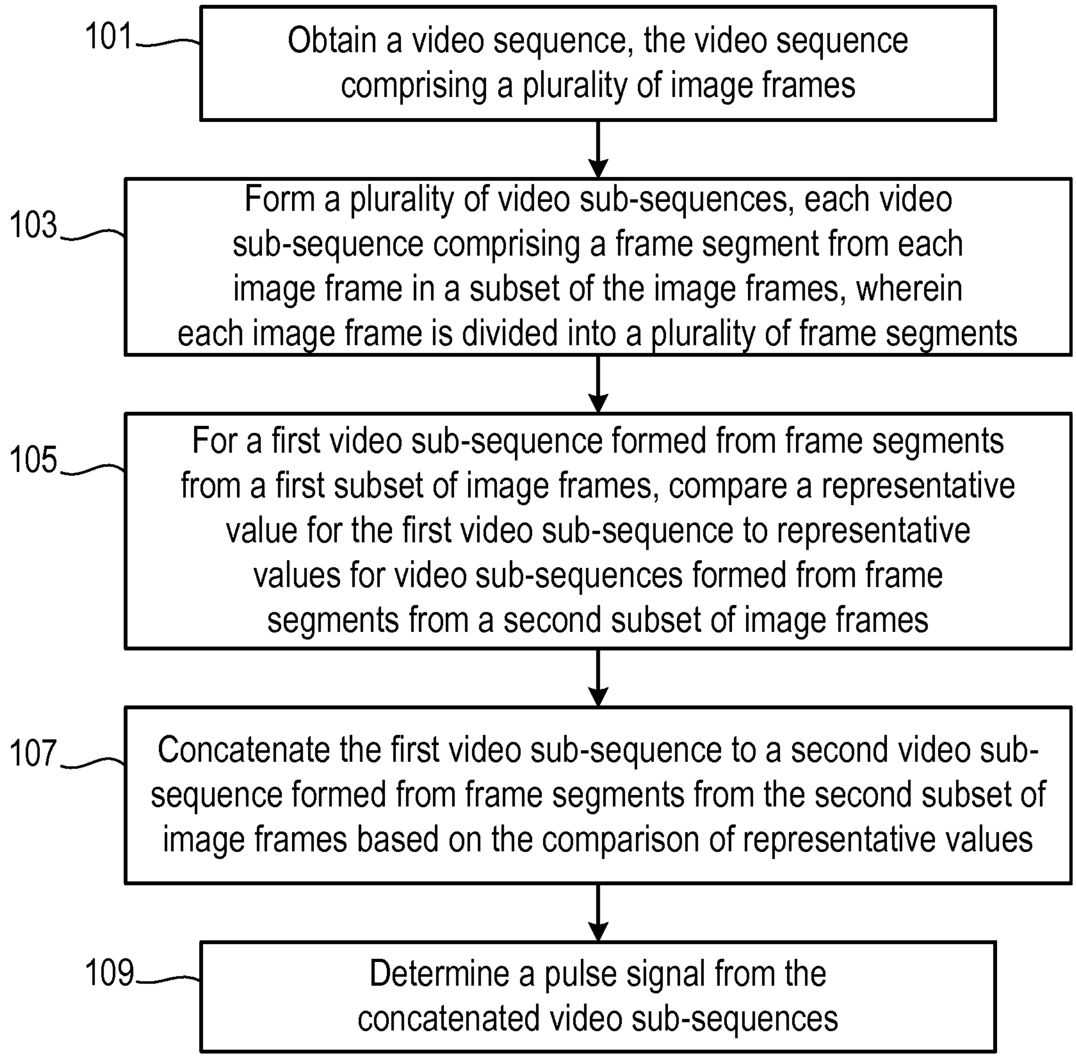
101
Obtain a video sequence, the video sequence comprising a plurality of image frames
103
Form a plurality of video sub-sequences, each video sub-sequence comprising a frame segment from each image frame in a subset of the image frames, wherein each image frame is divided into a plurality of frame segments
105
For a first video sub-sequence formed from frame segments from a first subset of image frames, compare a representative value for the first video sub-sequence to representative values for video sub-sequences formed from frame segments from a second subset of image frames
107
Concatenate the first video sub-sequence to a second video sub-sequence formed from frame segments from the second subset of image frames based on the comparison of representative values
109
Determine a pulse signal from the concatenated video sub-sequences

Figure 3

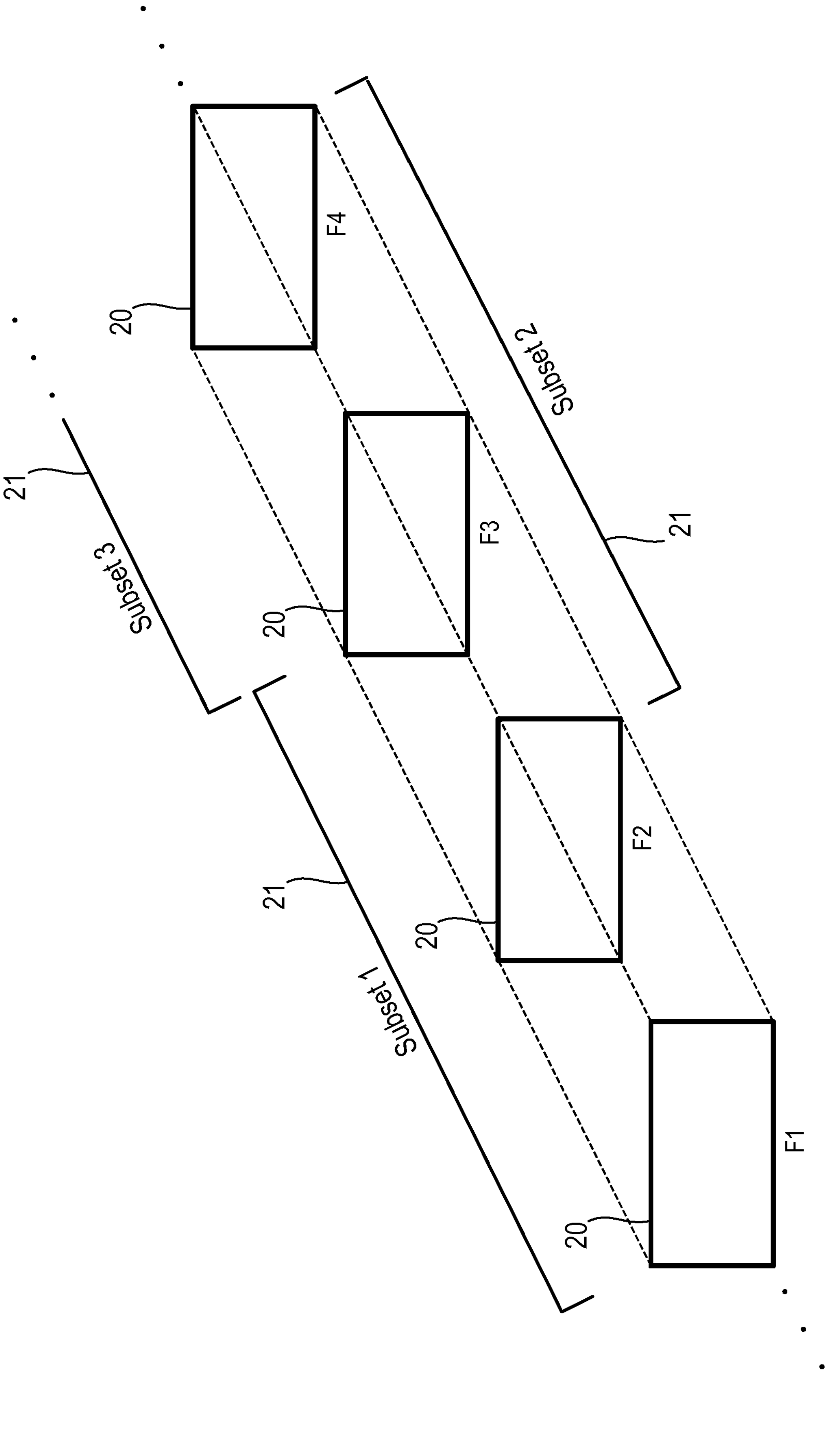

Figure 4(a)

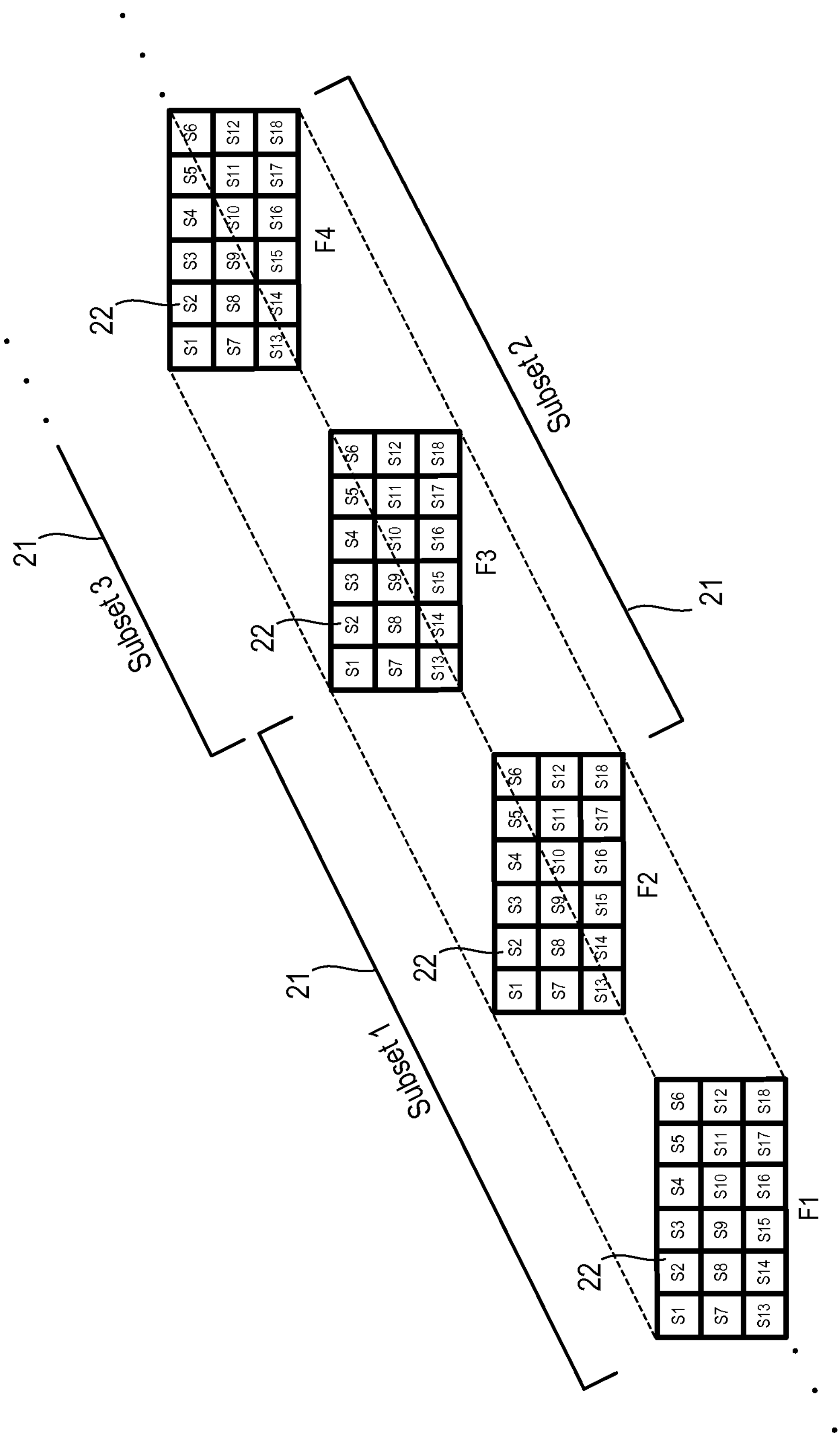
Subset 1
Subset 2
Subset 3
21
22
F1
F2
F3
F4
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18

Figure 4(b)

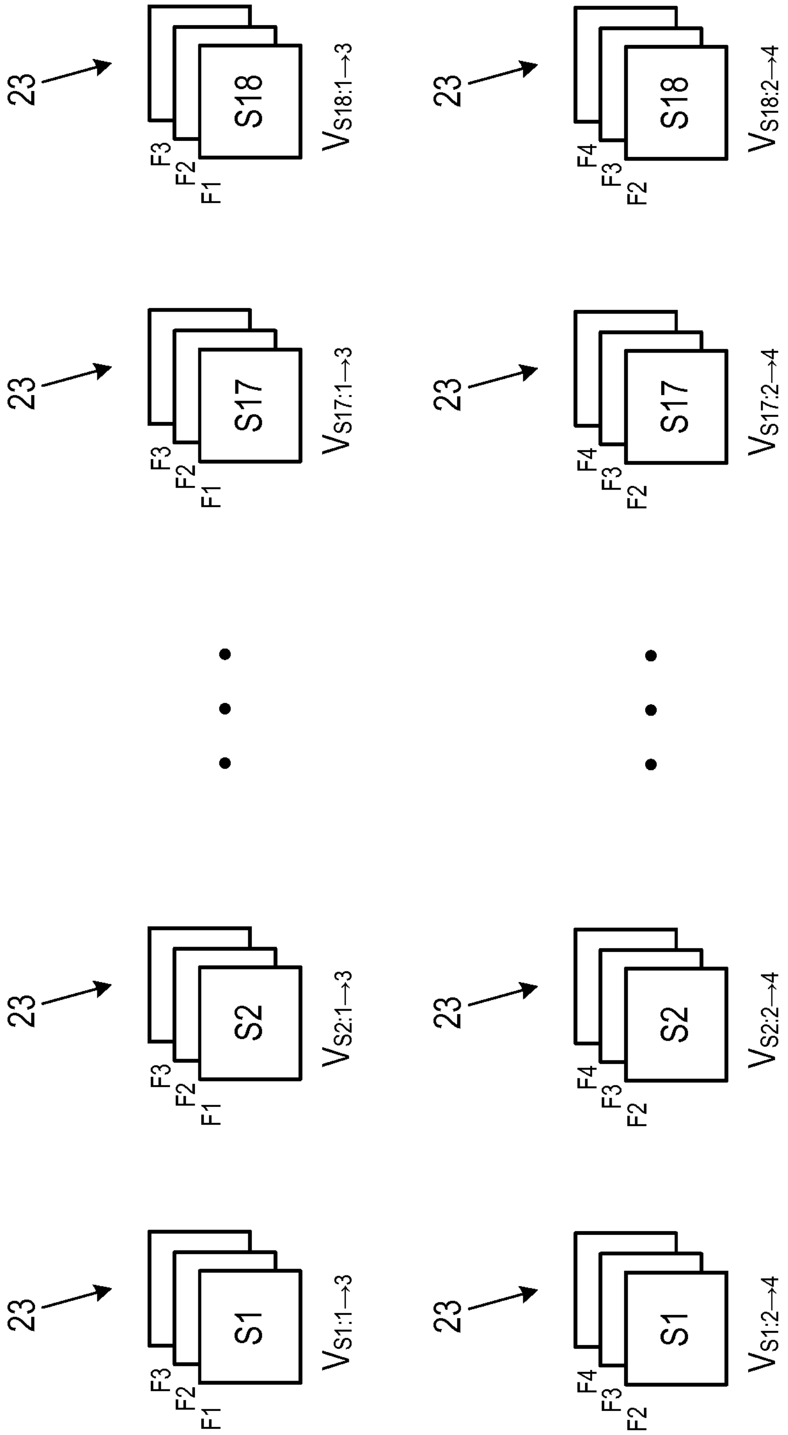
23
S1
S2
S17
S18
F1
F2
F3
F4
$V_{S1:1\rightarrow3}$
$V_{S2:1\rightarrow3}$
$V_{S17:1\rightarrow3}$
$V_{S18:1\rightarrow3}$
$V_{S1:2\rightarrow4}$
$V_{S2:2\rightarrow4}$
$V_{S17:2\rightarrow4}$
$V_{S18:2\rightarrow4}$

Figure 4(c)

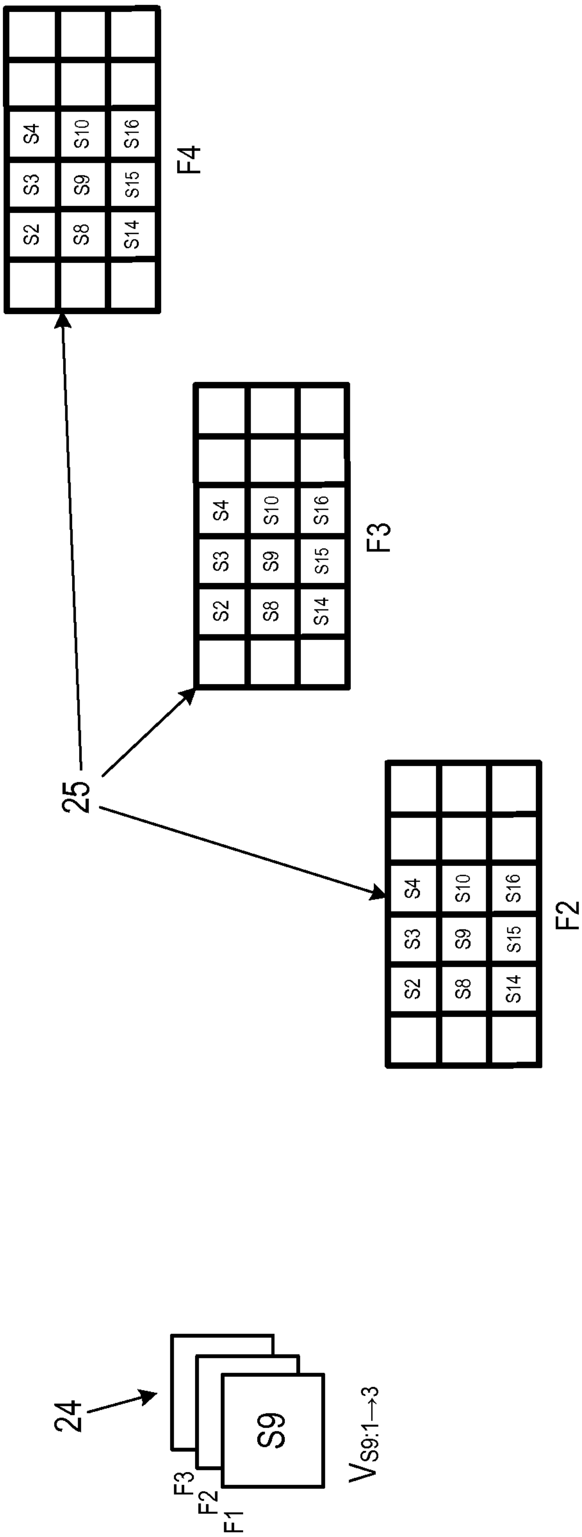

Figure 4(d)

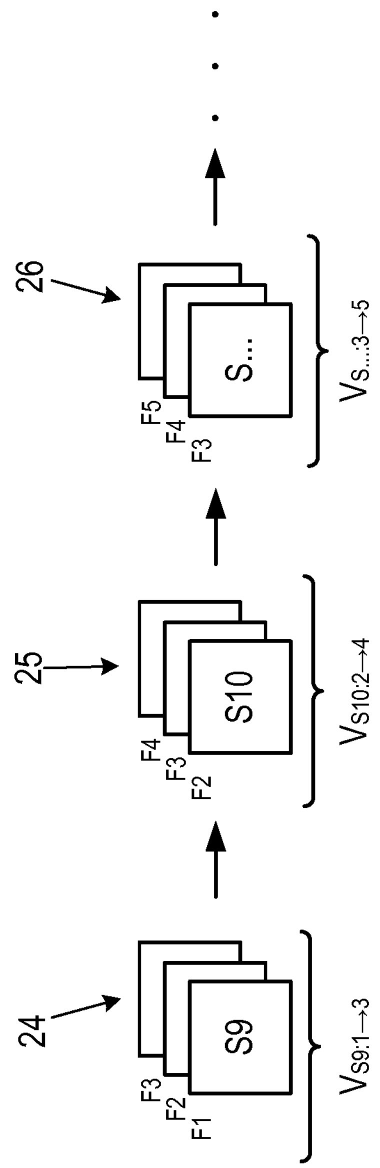

Figure 4(e)

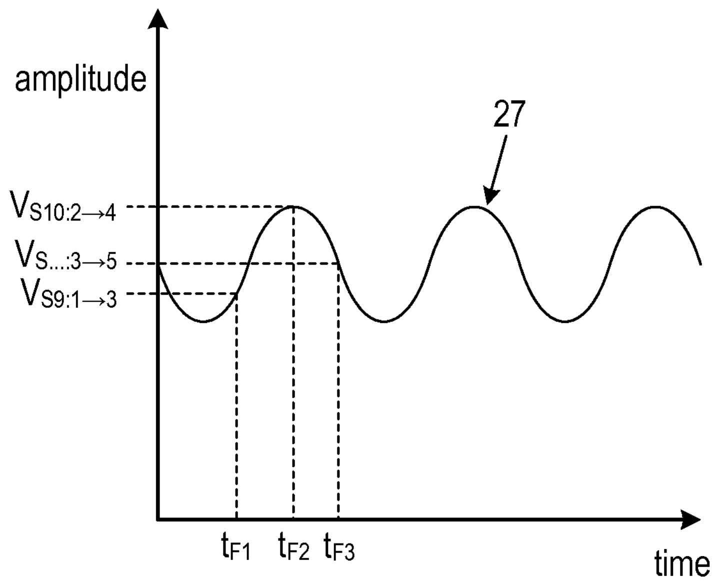
amplitude
27
$V_{S10:2\rightarrow4}$
$V_{S...:3\rightarrow5}$
$V_{S9:1\rightarrow3}$
t_{F1}
t_{F2}
t_{F3}
time

Figure 4(f)

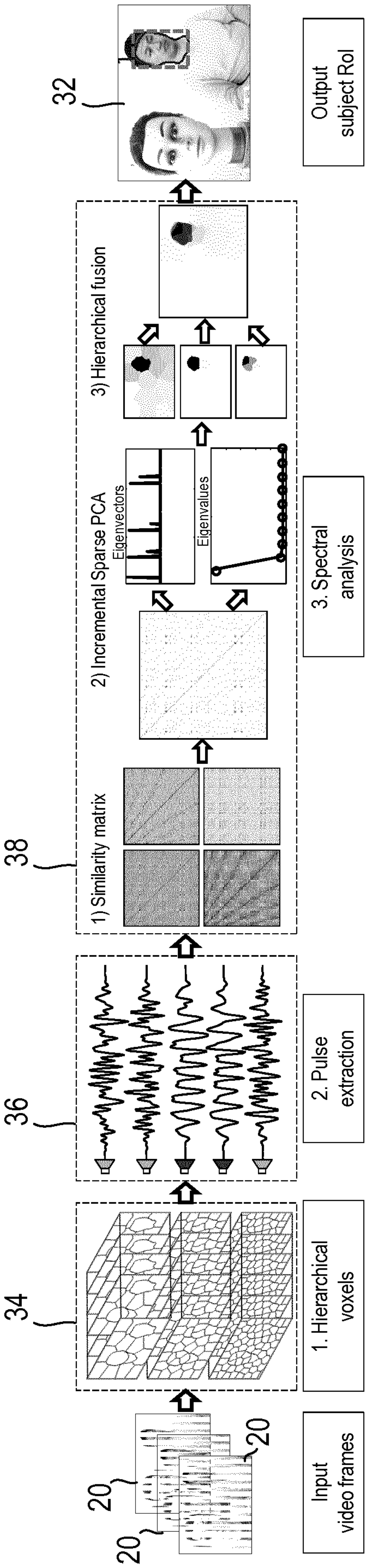
Input
video frames
20
20
20
1. Hierarchical
voxels
34
2. Pulse
extraction
36
38
1) Similarity matrix
2) Incremental Sparse PCA
Eigenvectors
Eigenvalues
3) Hierarchical fusion
3. Spectral
analysis
32
Output
subject RoI

Figure 5

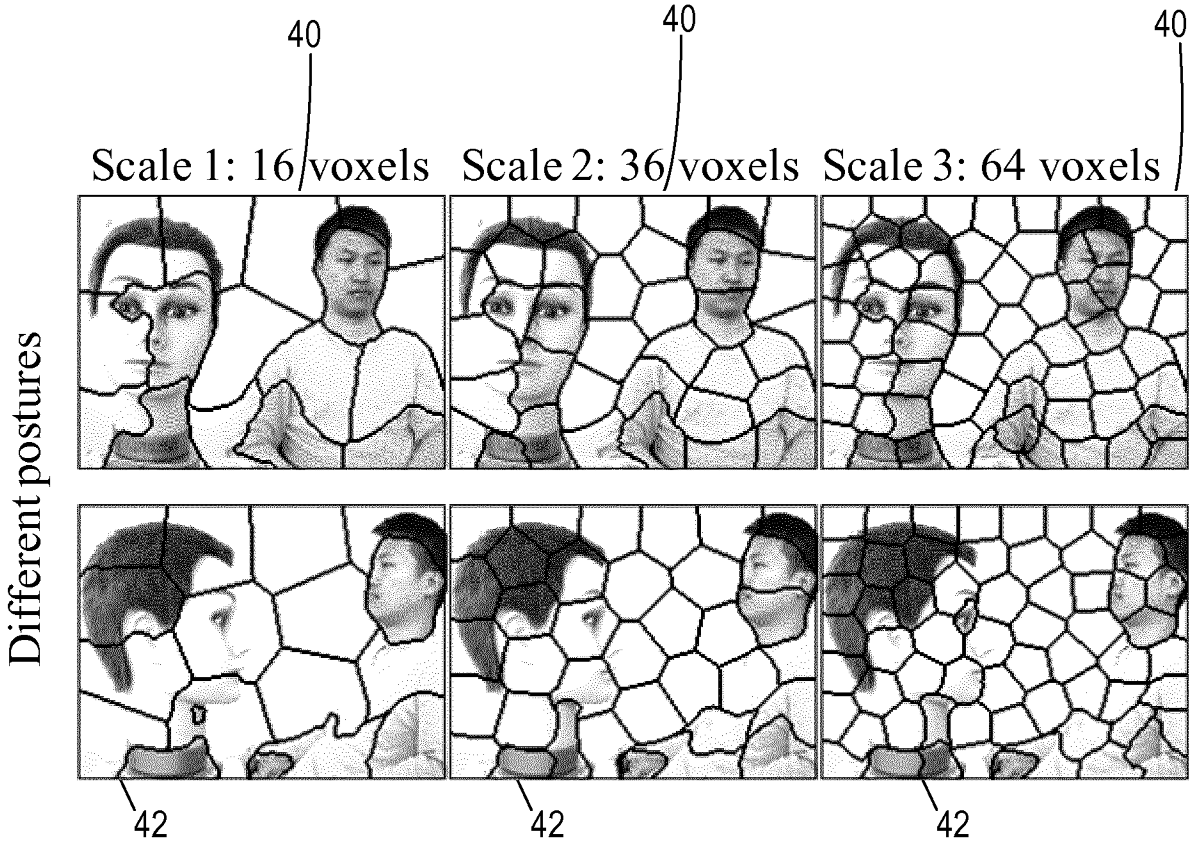
40
40
40
Scale 1: 16 voxels
Scale 2: 36 voxels
Scale 3: 64 voxels
Different postures
42
42
42

Figure 6

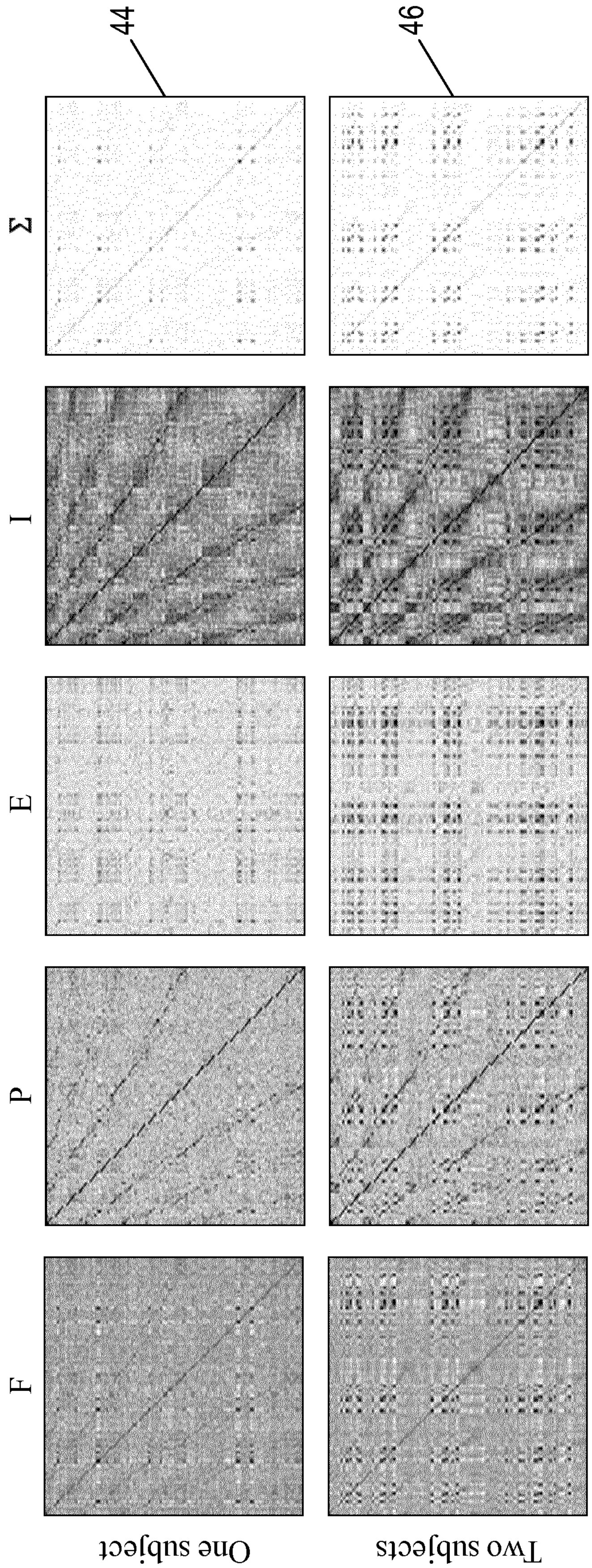

Figure 7

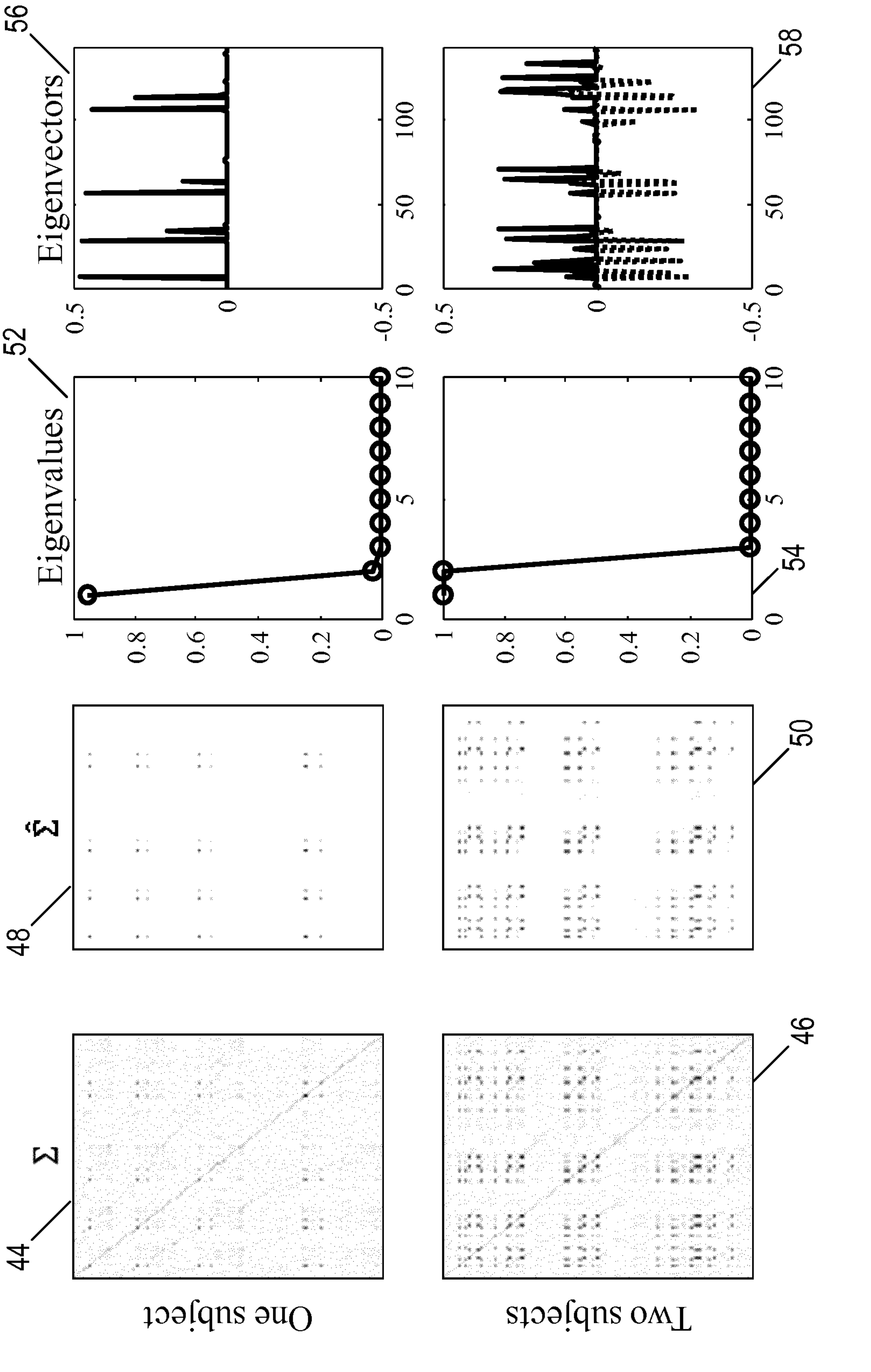
One subject
Two subjects
Σ
Σ̂
Eigenvalues
Eigenvectors
44
46
48
50
52
54
56
58

Figure 8

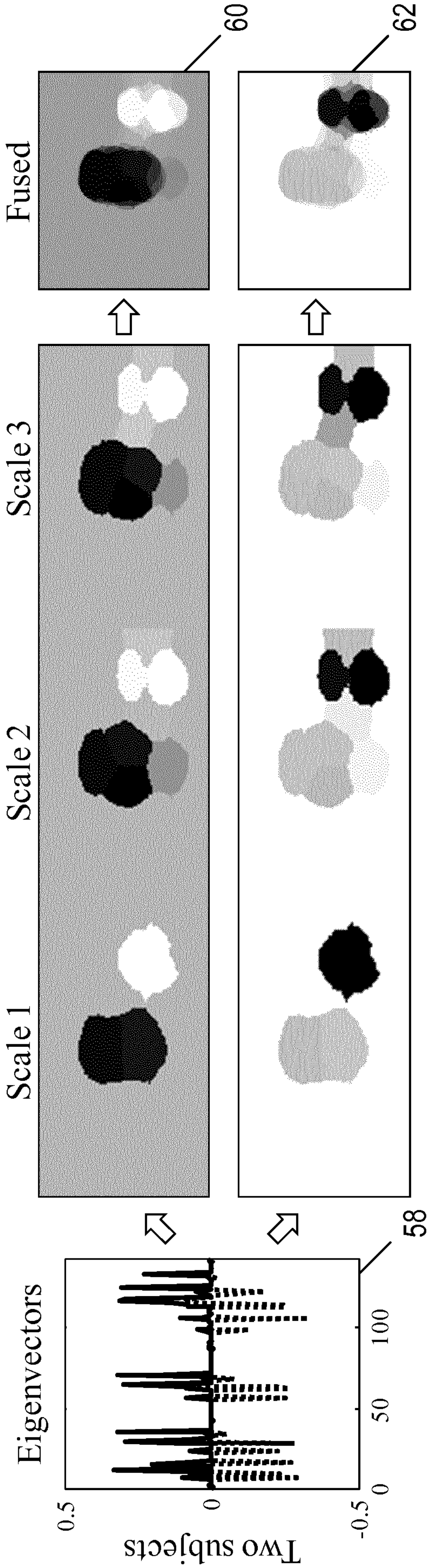
Eigenvectors
Two subjects
0.5
0
-0.5
0
50
100
58
Scale 1
Scale 2
Scale 3
Fused
60
62

Figure 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/060250

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/00 G06K9/62 G06T7/00 A61B5/024
ADD.

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)
G06K G06T A61B

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)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Ron Van Luijtelaar ET AL: "Automatic RoI Detection for Camera-Based Pulse-Rate Measurement" In: "LECTURE NOTES IN COMPUTER SCIENCE", 2 November 2014 (2014-11-02), Springer Berlin Heidelberg, Berlin, Heidelberg, XP055194698, ISSN: 0302-9743 ISBN: 978-3-54-045234-8 vol. 9009, pages 360-374, DOI: 10.1007/978-3-319-16631-5_27,	1-15
Y	sec. 2.1, par. 1-3, fig. 1; fig. 2; eqs. 2, 4, 7, 10-11; sec. 3.3, sentence 4; page 6, pars. 1-2, eq. 2, 8; page 9, sec. Cluster growing ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C.

[X] 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
23 June 2016	01/07/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Loza, Artur

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/060250

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Pratik Sahindrakar: "Improving Motion Robustness of Contact-less Monitoring of Heart Rate Using Video Analysis", , 24 August 2011 (2011-08-24), XP055051521, Eindhoven, NL Retrieved from the Internet: URL:http://alexandria.tue.nl/extra1/afstversl/wsk-i/sahindrakar2011.pdf [retrieved on 2013-01-29] sec. 3.6; sec. 3.7, par. 1; fig. 3.3; sec. 3.1, par. 1 and sec. 3.3 par. 1 -----	1-15
Y	WO 2011/042839 A1 (KONINKL PHILIPS ELECTRONICS NV [NL]; KIRENKO IHOR OLEHOVYCH [NL]; JEAN) 14 April 2011 (2011-04-14) page 16, line 8--page 17, line 10; , step 39; page 22, penultimate par., page 16, par. 4 -----	1-15
A	RESO MATTHIAS ET AL: "Temporally Consistent Superpixels", 2013 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, IEEE, 1 December 2013 (2013-12-01), pages 385-392, XP032572909, ISSN: 1550-5499, DOI: 10.1109/ICCV.2013.55 [retrieved on 2014-02-28] sec. 3, eqs. 1 and 2; page 388, Algorithm 1 -----	3-6,11, 12,14,15
A	SYLVAIN PARIS ET AL: "A gentle introduction to bilateral filtering and its applications", COMPUTER GRAPHICS PROCEEDINGS, PROCEEDINGS OF SIGGRAPH ANNUALINTERNATIONAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVETECHNIQUES, XX, XX, 5 August 2007 (2007-08-05), pages 1-45, XP002581515, eq. 3 -----	6,14,15
Y	WANG WENJIN ET AL: "Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE SERVICE CENTER, PISCATAWAY, NJ, USA, vol. 62, no. 2, 1 February 2015 (2015-02-01), pages 415-425, XP011570468, ISSN: 0018-9294, DOI: 10.1109/TBME.2014.2356291 [retrieved on 2015-01-16] section III.A.3; equations 1-3 -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2016/060250

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2011042839 A1	14-04-2011	CN 102576458 A	11-07-2012
		EP 2486539 A1	15-08-2012
		JP 5856960 B2	10-02-2016
		JP 2013506523 A	28-02-2013
		US 2012195486 A1	02-08-2012
		US 2014153800 A1	05-06-2014
		US 2015104088 A1	16-04-2015
		WO 2011042839 A1	14-04-2011