

(51) International Patent Classification:
G06K 9/46 (2006.01)

(21) International Application Number:

PCT/SG2014/000275

(22) International Filing Date:

12 June 2014 (12.06.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

201304548-9 12 June 2013 (12.06.2013) SG

(71) Applicant: **AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH** [SG/SG]; 1 Fusionopolis Way, #20-10 Connexis, Singapore 138632 (SG).(72) Inventors: **CHUA, Teck Wee**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (SG). **PHAM, Nam Trung**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (SG). **LEMAN, Karianto**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (SG). **CHANG, Richard Philippe**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (SG). **WANG, Yue**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (SG).(74) Agent: **SPRUSON & FERGUSON (ASIA) PTE LTD**; P.O. Box 1531, Robinson Road Post Office, Singapore 903031 (SG).

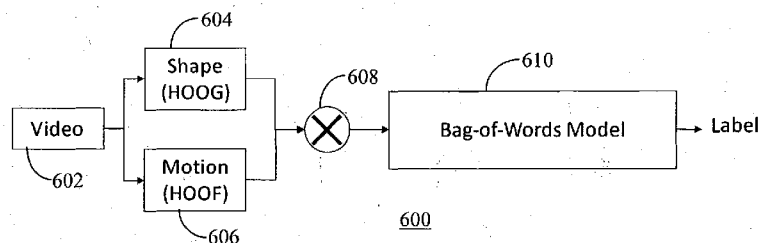
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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).

Published:

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

(54) Title: METHOD AND SYSTEM FOR HUMAN MOTION RECOGNITION

**FIG. 6**

(57) Abstract: A system and method for human motion recognition are provided. The system includes a video sequence decomposer, a feature extractor, and a motion recognition module. The video sequence decomposer decomposes a video sequence into a plurality of atomic actions. The feature extractor extracts features from each of the plurality of atomic actions, the features including at least a motion feature and a shape feature. And the motion recognition module performs motion recognition for each of the plurality of atomic actions in response to the features.



METHOD AND SYSTEM FOR HUMAN MOTION RECOGNITION

PRIORITY CLAIM

[0001] The present application claims priority to Singapore Patent Application No. 201304548-9, filed 12 June, 2013.

FIELD OF THE INVENTION

[0002] The present invention generally relates to surveillance and recognition technology, and more particularly relates to a system and method for human motion recognition.

BACKGROUND OF THE DISCLOSURE

[0003] There has been a surge, in recent years, towards the study of human action recognition because it is fundamental to many computer vision applications such as video surveillance, human-computer interface, and content-based video retrieval. While the human brain can recognize an action in a seemingly effortless fashion, recognition solutions using computers have, in many cases, proved to be immensely difficult.

[0004] One challenge is the choice of optimal representations for human actions. Ideally, the representation should be robust against inter- or intra-variations, noises, temporal variations, and sufficiently rich to differentiate a large number of possible actions. Practically, such representations do not exist.

[0005] It is well documented that human actions can be encoded as spatial information of body poses and dynamic information of body motions. However, some actions cannot be distinguished solely using shape and/or motion features. For

example, a skip action may look very similar to a run action if only the pose of the body is observed.

[0006] The classification task would be simplified if the motion flow of the entire body is considered simultaneously. Using this approach, one would expect that the skip action generates more vertical flows (upward and downward flows) than the run action. In addition, actions such as jogging, walking and running can be easily confused if only the pose information is used due to the similarity of postures in the action sequences.

[0007] Likewise, there are some actions which cannot be fully described by motion feature alone. Combining both motion and shape cues potentially provides complementary information about an action. Thus, conventionally, motion and shape feature vectors are concatenated to form a super vector. However, the super vector obtained through such concatenation may not explicitly convey the underlying action. Moreover, the super vector is unnecessarily long and requires complex feature dimension reduction techniques.

[0008] Thus, what is needed is a system and method for efficient recognition of human motion. Furthermore, other desirable features and characteristics will become apparent from the subsequent detailed description and the appended claims, taken in conjunction with the accompanying drawings and this background of the disclosure.

SUMMARY

[0009] According to the Detailed Description, a method for human motion recognition is provided. The method includes decomposing a video sequence into a plurality of atomic actions and extracting features from each of the plurality of atomic actions. The features extracted include at least a motion feature and a shape feature.

The method further includes performing motion recognition for each of the plurality of atomic actions in response to the features.

[0010] In accordance with another aspect, a system for human motion recognition is provided. The system includes a video sequence decomposer, a feature extractor, and a motion recognition module. The video sequence decomposer decomposes a video sequence into a plurality of atomic actions. The feature extractor extracts features from each of the plurality of atomic actions, the features including at least a motion feature and a shape feature. And the motion recognition module performs motion recognition for each of the plurality of atomic actions in response to the features.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying figures, where like reference numerals refer to identical or functionally similar elements throughout the separate views and which together with the detailed description below are incorporated in and form part of the specification, serve to illustrate various embodiments and to explain various principles and advantages in accordance with a present embodiment.

[0012] FIG. 1, comprising FIGs. 1A to 1D, pictorially illustrates human motion action sequences, wherein FIG. 1A illustrates action sequences of the human motion of walk, FIG. 1B illustrates action sequences of the human motion of run, FIG. 1C illustrates the human motion of skip, and FIG. 1D illustrates the human motion of run for comparison to the human motion of skip (FIG. 1C).

[0013] FIG. 2, comprising FIGs. 2A, 2B and 2C, depicts decomposition of human walking action in accordance with a present embodiment, wherein FIG. 2A depicts right-leg stepping, FIG. 2B two-leg crossing, and FIG. 2C left-leg stepping.

[0014] FIG. 3, comprising FIGs. 3A, 3B and 3C, depicts motion flow and shape flow vectors of the human motion of running in accordance with the present embodiment, wherein FIG. 3A depicts the motion of running with both vectors, FIG. 3B depicts the motion flow vectors only, and FIG. 3C depicts the shape flow vectors only.

[0015] FIG. 4, comprising FIGs. 4A and 4B, depicts illustrations of histogram binning in accordance with the present embodiment, wherein FIG. 4A depicts a graph of the human motion vectors of FIG. 3 and FIG. 4B depicts histogram binning the vectors of FIG. 4A in accordance with the present embodiment.

[0016] FIG. 5, comprising FIGs. 5A and 5B, depicts regional concatenation histogram binning in accordance with the present embodiment, wherein FIG. 5A depicts a bounding box illustrating a human motions and divided into four regions and FIG. 5B depicts concatenation of resultant histograms from each region.

[0017] FIG. 6 depicts a block diagram of a system for resolving an action video into a collection of repetitive atomic actions in accordance with the present embodiment.

[0018] FIG. 7 depicts a representation of a first portion of the operation of the system of FIG. 6 where the action video of human motion on the left is resolved into a combination of regional concatenation of resultant histograms in accordance with the present embodiment

[0019] FIG. 8 depicts a representation of a second portion of the operation of the system of FIG. 6 where the resultant histograms are clustered into distance weighted bag-of-automatic-actions in accordance with the present embodiment.

[0020] FIG. 9, comprising FIGs. 9A, 9B and 9C, depicts motion recognition in accordance with the present embodiment of six different human action sequences from two different human action sequence datasets, wherein FIG. 9A illustrates the

six different human action sequences utilized for the comparison, FIG. 9B is a graph of the accuracy of recognition of various human action features in a first human action sequence dataset by motion recognition in accordance with the present embodiment, and FIG. 9C is a graph of the accuracy of recognition of the various human action features in a second human action sequence dataset by motion recognition in accordance with the present embodiment.

[0021] And FIG. 10 illustrates a 6x6 matrix of the motion recognition in accordance with the present embodiment of the six different human action sequences in both human action sequence datasets of FIG. 9 where the six different human action sequences are listed along the vertical and the recognized human actions are listed along the horizontal.

[0022] Skilled artisans will appreciate that elements in the figures are illustrated for simplicity and clarity and have not necessarily been depicted to scale. For example, the dimensions of some of the elements in the block diagrams or flowcharts may be exaggerated in respect to other elements to help to improve understanding of the present embodiments.

DETAILED DESCRIPTION

[0023] The following detailed description is merely exemplary in nature and is not intended to limit the invention or the application and uses of the invention. Furthermore, there is no intention to be bound by any theory presented in the preceding background of the invention or the following detailed description. It is the intent of this invention to present an efficient and recognition of human action with improved accuracy.

[0024] As stated above, human actions can be encoded as spatial information of body poses and dynamic information of body motions. Referring to FIG. 1,

comprising FIGs. 1A to 1D, action sequences of human actions are shown. In FIG. 1A, an action sequence 100 depicts five views 102, 104, 106, 108, 110 of a person walking. In FIG. 1B, an action sequence 120 depicts five views 122, 124, 126, 128, 130 of a person running. Some actions cannot be distinguished solely using shape and/or motion features. In FIGs. 1C and 1D, two views 140, 150 depict single body poses of two different human actions. It is difficult to distinguish a skip action in the view 140 from a run action in the view 150 because when only the pose of the body is observed at a single point in the action, the skip action view 140 appears very similar to the run action view 150.

[0025] In accordance with the present embodiment, a complex human action sequence is decomposed into a sequence of elementary building blocks, known as 'atomic actions'. Referring to FIG. 2, comprising FIGs. 2A, 2B and 2C, the human action of walking is broken down into several atomic actions 200, 210, 220. FIG. 2A depicts the atomic action 200 of right-leg stepping, FIG. 2B depicts the atomic action 210 of two-leg crossing, and FIG. 2C depicts the atomic action 220 left-leg crossing.

[0026] Shape and motion are the two most important cues for actions, and atomic actions can be 'synthesized' from both elements. FIG. 3, comprising FIGs. 3A, 3B and 3C, illustrates these cues. In the view 300 of FIG. 3A an atomic action of a person 302 running is shown. A first set of vectors 304 are vectors depicting motion flow (e.g., right forearm moving up, left forearm moving forward, left leg moving forward, right lower leg moving up). The first set of vectors 304 are extracted into a motion flow vector view 310 in FIG. 3B. A second set of vectors 306 are vectors depicting shape flow (i.e., vectors identifying the outline of the shape). The second set of vectors 306 are extracted into a shape flow vector view 320 in FIG. 3B.

[0027] Observing shape and motion is a very natural way to recognize an action. The visual cortex in the brain has two pathways to process shape and motion information. Motivated by the robustness of histograms of features, in accordance with a present embodiment a histogram-of-oriented gradient (HOOG) and a histogram-of-oriented optical flow (HOOF) are used as shape and motion descriptors, respectively. HOOG is also be used as a pose descriptor.

[0028] Such a feature is more robust against scale variation and the change of motion direction. A method for extraction of the HOOF and the HOOG in accordance with the present embodiment is illustrated in FIG. 4, comprising FIGs. 4A and 4B. In FIG. 4A, a graph 400 of human motion vectors includes a horizontal axis 405 and a vertical axis 410. Pose and shape vectors 412, 414, 416, 418 plotted on the graph 400 and are divided in accordance with their primary angles from the horizontal axis 405 (i.e., their symmetry about the vertical axis 410).

[0029] Referring to FIG. 4B, a four-bin histogram 450 bins the vectors 412, 414, 416, 418 in accordance with their symmetry about the vertical axis 410—that is the vectors 412 are binned in bin 452, the vectors 414 are binned in bin 454, the vectors 416 are binned in bin 456, and the vectors 418 are binned in bin 458. In this manner, the histogram bar graph 450 depicts human motion feature extraction of pose, shape and motion as a histogram combination of the HOOG and the HOOF. The histogram 450 is normalized to sum up to unity to make it scale-invariant. In this manner, there is no need to normalize the size of each of the bounding boxes 452, 454, 456, 458.

[0030] As a result, the histogram of a person moving from left to right will be the same as a histogram of a person moving from right to left (i.e., the method in accordance with the present embodiment is direction indiscriminate). The

contribution of each vector is proportional to its magnitude and the histogram is normalized to sum up to unity to make it scale-invariant.

[0031] FIG. 5, comprising FIGs. 5A and 5B, depicts regional concatenation histogram binning in accordance with the present embodiment. Spatial information is considered by dividing a bounding box 500 of a subject 510 into a 4×4 matrix of regions 512, 514, 516, 518, as shown in FIG. 5A. A resultant histogram 550 depicts concatenation of four-bin histograms 552, 554, 556, 558 from each region.

[0032] As discussed above in regards to FIG. 2, a complex action can be decomposed into a sequence of elementary building blocks known as atomic actions. Referring to FIG. 6, a block diagram 600 depicts a system in accordance with the present embodiment. An atomic action is defined as the action performed at video frame t of a video 602. It is represented by a shape histogram (i.e., HOOG) 604 extracted at frame t and an optical flow histogram (i.e., HOOF) 606 computed at frames $(t-1)$ and t . Therefore, the T -frame action video 602 has $(T - 1)$ number of atomic actions.

[0033] Suppose action, shape, and motion are three discrete random variables: Z ; S ; M with distribution $z[x]$; $s[x]$; and $m[x]$ respectively, where $[]$ represents discrete data. $s[x]$ and $m[x]$ are basically the shape and motion histograms computed. In a further assumption, an action is a function of shape and motion, i.e., $Z = f(S;M)$. The simplest function would be a summation:

$$Z = S + M \quad (1)$$

[0034] According to probability theory, the sum of discrete random variables will produce a new random variable with distribution that can be determined via convolution. Therefore, the distribution (histogram) of an action can be determined by

$$z[x] = s[x] * m[x] = \sum_{k=-\infty}^{k=+\infty} s[k].m[x-k] \quad (2)$$

where the asterisk ‘*’ denotes the convolution operator. The idea of using a convolution operator is also inspired by success of convolution-based reverb applications in digital signal processing (DSP). In DSP, convolution is a mathematical way of combining two source signals to form an output signal. The output signal bears the characteristics of both sources. Convolution-based reverb is a process for digitally simulating the reverberation of a virtual or physical space. Given the impulse response of a space which can be obtained by recording a short burst of a broadband signal, any “dry” signal (little room or space influence) can be convolved with the impulse response. The result is that the sound appears to have been recorded in that space. Analogously, knowing that an action is characterized by both shape and motion information, an atomic action histogram can be obtained by convolving the corresponding shape histogram (HOOG) 604 and motion histogram (HOOF) 606. The convolution operation 608 is commutative, which means that the order of the inputs does not mathematically matter.

[0035] The length of the output is given by the expression $\|s\| + \|m\| - 1$. This representation has two major advantages. First, the action histogram is more robust against noises. This is because each bin in the action histogram is influenced by bins in the shape histogram weighted by the motion histogram or vice versa (the commutative property of convolution). Therefore the effect of abrupt changes in the histogram magnitude can be minimized. Second, the action histogram produced using convolution is more discriminative. The ratio of inter-class distance to intra-class distance is measured and the results on a known human action video dataset is shown below in Table 1.

[0036] Table 1 shows a comparison of normalized inter-/intra-class distance ratio on a known human action video dataset for different types of feature combination methods where a Hellinger distance measure is used to compare two histograms:

$$D_h(s[x], m[x]) = \left[1 - \sum_{\forall x} \sqrt{s[x] \cdot m[x]} \right]^{\frac{1}{2}} \quad (3)$$

A higher value indicates that the feature is potentially more discriminative.

Combination Strategies	Ratio
Convolution (<i>Conv</i>)	1.0000
Summation (<i>Sum</i>)	0.8535
Product (<i>Prod</i>)	0.8489
Concatenation (<i>Concat</i>)	0.8743

Table 1

The results suggest that the convolution operation produces feature vectors that are potentially more discriminative than the features obtained through other combination methods.

[0037] In one example, an action video is represented as a collection of repetitive atomic actions. The basic concept is illustrated in FIG. 6. Recall that atomic actions are characterized by the convolved shape-motion histogram. A visual codebook can be created by performing K-means clustering on all atomic actions from the training data. The cluster centroids serve as the visual codewords. Next, each atomic action in the video is compared against those codewords and the distances are recorded accordingly. The distance between the atomic action and its nearest codeword is used to weight the histogram bin. The histogram for all relevant codewords in a video is computed by aggregating their respective distances. This final representation allows

any lengthy video to be 'compressed' into a compact histogram. The histogram is normalized such that a sum of the bins is unity. The normalization ensures that the histogram distribution is invariant to the video length. For instant, given a particular action class, codewords (i.e., key atomic actions) frequencies for a variable length video remain relatively stable. The aforementioned compact video representation can also be called distance weighted bag-of-automatic-actions (i.e., bag-of-words model 610).

[0038] Referring to FIG. 7, another representation 700 of the system 600 (FIG. 6) takes an atomic action 702 and extracts a HOOG histogram 704 and a HOOF histogram 706. The histograms 704, 706 are convoluted to form a distance weighted bag-of-automatic-actions resultant histogram 708.

[0039] Referring to FIG. 8, a representation 800 of a second portion of the operation of the system 600 where the resultant histograms are clustered into distance weighted bag-of-automatic-actions in accordance with the present embodiment is depicted. A human action is broken down into several atomic actions 802. Histograms of the atomic actions 802 are clustered with a K-means operation as shown in a graph 804 of the clusters. Then, each atomic action 802 is assigned to its closest cluster center 806, 808, 810. In this manner, a contribution of each atomic action 802 to a histogram 820 is inversely proportional to the atomic action's distance from a cluster center. The entire video sequence is then represented in the distance weighted occurrence histogram 820 of visual words in bins 822, 824, 826, 828, 830. At a normalization step 840, a sum of the bins 822, 824, 826, 828, 830 is normalized to one. Thus, the sum of the bins 822, 824, 826, 828, 830 is normalized to one is unity and, thus, is advantageously invariant to a video length of the entire video sequence of the atomic actions 802.

[0040] The action recognition framework in accordance with the present embodiment has been evaluated using a first and a second publicly available dataset, identified as the Weizmann dataset (the first dataset) and the KTH dataset (the second dataset). The KTH dataset has been regarded either as one large set with strong intra-subject variations (all-in-one) or as four independent scenarios. In the latter case, each scenario is trained and tested separately. For the KTH-based evaluation, the focus was on the KTH all-in-one case.

[0041] Since the KTH dataset size is much larger than the Weizmann dataset size, a K-means algorithm is used to cluster the training data as seen in the graph 804. Each class in the KTH dataset is quantized into five hundred clusters. This quantization can reduce the intra-class variation and computational time. A leave-one-out cross validation (LOOCV) protocol is used in all of the evaluations. Table 2 shows the LOOCV recognition rate.

Table 2. LOOCV classification accuracy using different number of clusters.

Dataset	No. of Clusters	Convol	Sum	Prod	Concat	HOOG	HOOF
Weizmann	5	96.67	87.78	87.78	86.67	66.67	73.33
	10	100	94.44	94.44	93.33	74.44	78.89
	15	100	97.78	94.44	98.89	77.78	81.11
	20	100	97.78	96.67	97.78	78.89	84.44
	25	100	98.89	96.67	97.78	81.11	83.33
KTH	10	83.94	70.25	72.22	75.58	45.88	57.73
	25	91.63	79.94	79.92	83.30	51.90	63.88
	40	92.46	82.44	81.43	87.64	45.24	64.88
	55	91.46	84.43	83.62	86.97	45.58	63.37

Table 2

For the Weizmann dataset which only uses five clusters (codewords), the convolved feature yields a much higher accuracy (96.67%) as compared to other features. When the number of clusters is increased further, the convolved feature consistently gives perfect classification accuracy (100%). Using only shape feature (HOOG) or only motion features (HOOF) results in poorer results than using a method in accordance

with the present embodiment. On average, the method in accordance with the present embodiment provided about 11.29% overall improvement as compared to other methods.

[0042] Referring to FIG. 9, comprising FIGs. 9A, 9B and 9C, motion recognition in accordance with the present embodiment of human action sequences 900 including six different human action sequences 902, 904, 906, 908, 910, 912 from the KTH human action sequence datasets is performed, examined and compared. In FIG. 9A, the six different human action sequences utilized for the comparison are boxing 902, clapping 904, waving 906, walking 908, jogging 910, and running 912. FIG. 9B is a bar graph 920 of the accuracy (plotted along a vertical axis 924) of recognition of various human action features using different numbers of clusters (plotted along a horizontal axis 922) in the Weizmann human action sequence dataset by motion recognition in accordance with the present embodiment using a variety of K-means clustering operations (as described in accordance with FIG. 8). It can be seen that convolving histograms in accordance with the present embodiment provides the highest accuracy no matter how many clusters are used in the K-means clustering operations. FIG. 9C is a graph 940 of the accuracy (plotted along a vertical axis 944) of recognition of various human action features using different numbers of clusters (plotted along a horizontal axis 942) in the KTH human action sequence dataset by motion recognition in accordance with the present embodiment, also using a variety of K-means clustering operations. It can be seen that convolving histograms in accordance with the present embodiment as shown in the graph 940 provides the highest accuracy using the KTH dataset also, no matter how many clusters are used in the K-means clustering operations.

[0043] Higher accuracies are attained from the convolved feature for all number of clusters of the KTH dataset. The advantage of using the convolved feature is more prominent in the KTH dataset. The average improvement over all other five features is 19.56%. Again, the HOOG feature alone or the HOOF feature alone fails to provide discriminative information. One important observation from the results in the graphs 920, 940 is that the method and system in accordance with the present embodiment consistently requires a much smaller number of clusters or codewords to give higher accuracy. For example, with only ten clusters, operation in accordance with the present embodiment achieves comparable accuracy with a product feature which uses forty clusters. This confirms the finding that the convolved feature is significantly more discriminative than conventional features.

[0044] Referring to FIG. 10, results of the operations of FIG. 9 are summarized in a 6x6 matrix 1000 of the motion recognition in accordance with the present embodiment of the six different human action sequences 902, 904, 906, 908, 910, 912 in the KTH human action sequence dataset. The six different human action sequences are listed along the vertical 1002 and the recognized human actions are listed along the horizontal 1004. It can be seen that human motion recognition for walking, jogging and running are in the 90th percentile with the only misclassifications in these same three actions. Classifications for handclapping are also in the 90th percentile with the only misclassifications falling in handwaving and boxing. While misclassifications for boxing and handwaving are more plentiful, some of these misclassifications may be due to the KTH human action sequence dataset in that some of the atomic actions extracted may include an erroneous bounding box which is extracted off-centered from a body axis of the subject in the bounding box.

[0045] In the various examples illustrated above, a method to encode human actions by convolving shape-motion histograms has been presented. The main idea is to produce an output signal (i.e., an action histogram) from the source signals (i.e., shape and motion histograms) so that the output shares the characteristics of both source signals and inputs. The features are also much more discriminative than other hybrid features obtained through other combination strategies such as concatenation, sum, and product. Further, combination of shape and motion features greatly improves the classification results.

[0046] In addition, operation in accordance with the present embodiment avoids the need to determine weights manually during feature concatenation. The convolved feature is also very compact and has much lower dimensionality (79-dimensional) as compared to conventional concatenated features of 512-dimensional and 1000-dimensional methodologies. Due to the discriminative nature of the convolution feature, the codebook size is extremely small as compared to conventional methods. Also, the entire video sequence is advantageously represented as a distance weighted occurrence histogram of visual words.

[0047] Thus, it can be seen that a system and method for human motion recognition has been provided. The system includes a video sequence decomposer 602, a feature extractor (including HOOG 604 and HOOF 606), and a motion recognition module (including combiner 608 and bag-of-words model module 610). The video sequence decomposer decomposes a video sequence into a plurality of atomic actions. The feature extractor extracts features from each of the plurality of atomic actions, the features including at least a motion feature and a shape feature. And the motion recognition module performs motion recognition for each of the plurality of atomic actions in response to the features.

[0048] The motion recognition module performs motion recognition for each of the plurality of atomic actions by convolving histograms of the features of each of the plurality of atomic actions. In regards to the shape feature, the feature extractor extracts a set of shape vectors depicting shape flow from each of the plurality of atomic actions and the motion recognition module convolves histograms of the shape features of each of the plurality of atomic actions by deriving a shape descriptor by determining a histogram-of-oriented gradient of the set of shape vectors for each of the plurality of atomic actions.

[0049] In regards to the motion feature, the feature extractor extracts a set of motion vectors depicting motion flow from each of the plurality of atomic actions and the motion recognition module convolves histograms of the motion features of each of the plurality of atomic actions by deriving a motion descriptor by determining a histogram-of-oriented optical flow of the set of motion vectors for each of the plurality of atomic actions.

[0050] The features may also include a pose feature, and the feature extractor further extracts a set of pose vectors from each of the plurality of atomic actions and the motion recognition module convolves histograms of the pose features of each of the plurality of atomic actions by deriving a pose descriptor by determining a histogram-of-oriented gradient of the set of pose vectors for each of the plurality of atomic actions. The features may also include a spatial feature, and the feature extractor derives each of a set of shape, motion or pose vectors for each of two or more regions of a bounding-box within each of the plurality of atomic actions. The bounding box in each of the plurality of atomic actions is configured to include all of a subject pictured in the one of the plurality of atomic actions. The motion recognition module

convolves histograms of each of the shape, motion or pose descriptors to generate a resultant histogram.

[0051] The motion recognition module also normalizes the histograms of each of the plurality of atomic actions to sum up to unity and further may include a bag-of-words model module for K-means clustering of all of the atomic actions to generate a distance weighted bag-of-automatic-actions model of the video sequence.

[0052] Thus, in accordance with the present embodiment an efficient human motion recognition system and method is provided. The present embodiment is computationally efficient as compared to conventional motion recognition systems and even in comparison to conventional combination strategies such as sum, product and concatenation. The technology of the present embodiment and its various alternates and variants can be used for many scenarios. For example, the present embodiment provides a computationally efficient system and method for many computer vision applications such as video surveillance, human-computer interface, and content-based video retrieval which is robust against inter- or intra-variations, noises, temporal variations, and sufficiently rich to differentiate a large number of possible actions.

[0053] Thus, it can be seen that a system and method for human motion recognition which reduces complexity of the recognition methodology has been provided. While exemplary embodiments have been presented in the foregoing detailed description of the invention, it should be appreciated that a vast number of variations exist.

[0054] It should further be appreciated that the exemplary embodiments are only examples, and are not intended to limit the scope, applicability, operation, or configuration of the invention in any way. Rather, the foregoing detailed description will provide those skilled in the art with a convenient road map for implementing an

exemplary embodiment of the invention, it being understood that various changes may be made in the function and arrangement of elements and method of operation described in an exemplary embodiment without departing from the scope of the invention as set forth in the appended claims.

CLAIMS

What is claimed is:

1. A method for human motion recognition comprising:
decomposing a video sequence into a plurality of atomic actions;
extracting features from each of the plurality of atomic actions, the features comprising at least a motion feature and a shape feature; and
performing motion recognition for each of the plurality of atomic actions in response to the features.
2. The method in accordance with Claim 1 wherein the step of performing motion recognition for each of the plurality of atomic actions comprises performing motion recognition for each of the plurality of atomic actions by convolving histograms of the features of each of the plurality of atomic actions.
3. The method in accordance with Claim 2 wherein the step of extracting features from each of the plurality of atomic actions comprises extracting a set of shape vectors depicting shape flow from each of the plurality of atomic actions.
4. The method in accordance with Claim 3 wherein convolving histograms of the features of each of the plurality of atomic actions comprises deriving a shape descriptor by determining a histogram-of-oriented gradient of the set of shape vectors for each of the plurality of atomic actions.

5. The method in accordance with Claim 2 wherein the step of extracting features from each of the plurality of atomic actions comprises extracting a set of motion vectors depicting motion flow from each of the plurality of atomic actions.

6. The method in accordance with Claim 5 wherein convolving histograms of the features of each of the plurality of atomic actions comprises deriving a motion descriptor by determining a histogram-of-oriented optical flow of the set of motion vectors for each of the plurality of atomic actions.

7. The method in accordance with any of Claims 2 to 6 wherein the features further comprise a pose feature, and wherein the step of extracting features from each of the plurality of atomic actions comprises extracting a set of pose vectors from each of the plurality of atomic actions, and wherein convolving histograms of the features of each of the plurality of atomic actions comprises deriving a pose descriptor by determining a histogram-of-oriented gradient of the set of shape vectors for each of the plurality of atomic actions.

8. The method in accordance with any of the preceding claims wherein the features further comprise a spatial feature, and wherein the step of extracting features from each of the plurality of atomic actions comprises deriving each of a set of shape, motion or pose vectors for each of two or more regions of a bounding box within each of the plurality of atomic actions, the bounding box in each of the plurality of atomic actions configured to include all of a subject pictured in the one of the plurality of atomic actions, and wherein the step of performing motion recognition

for each of the plurality of atomic actions comprises convolving histograms of each of the shape, motion or pose descriptors to generate a resultant histogram.

9. The method in accordance with any of Claims 2 to 8 wherein convolving histograms of the features of each of the plurality of atomic actions comprises normalization of the histograms of each of the plurality of atomic actions to sum up to unity.

10. The method in accordance with any of the preceding claims further comprising K-means clustering of all of the atomic actions to generate a distance weighted bag-of-automatic-actions model of the video sequence.

11. A system for human motion recognition comprising:
a video sequence decomposer for decomposing a video sequence into a plurality of atomic actions;
a feature extractor for extracting features from each of the plurality of atomic actions, the features comprising at least a motion feature and a shape feature; and
a motion recognition module for performing motion recognition for each of the plurality of atomic actions in response to the features.

12. The system in accordance with Claim 11 wherein the motion recognition module performs motion recognition for each of the plurality of atomic actions by convolving histograms of the features of each of the plurality of atomic actions.

13. The system in accordance with Claim 12 wherein the feature extractor extracts a set of shape vectors depicting shape flow from each of the plurality of atomic actions.

14. The system in accordance with Claim 13 wherein the motion recognition module convolves histograms of the shape features of each of the plurality of atomic actions by deriving a shape descriptor by determining a histogram-of-oriented gradient of the set of shape vectors for each of the plurality of atomic actions.

15. The system in accordance with Claim 12 wherein the feature extractor extracts a set of motion vectors depicting motion flow from each of the plurality of atomic actions.

16. The system in accordance with Claim 15 wherein the motion recognition module convolves histograms of the motion features of each of the plurality of atomic actions by deriving a motion descriptor by determining a histogram-of-oriented optical flow of the set of motion vectors for each of the plurality of atomic actions.

17. The system in accordance with any of Claims 12 to 16 wherein the features further comprise a pose feature, and wherein the feature extractor further extracts a set of pose vectors from each of the plurality of atomic actions, and wherein the motion recognition module convolves histograms of the pose features of each of the plurality of atomic actions by deriving a pose descriptor by determining a

histogram-of-oriented gradient of the set of pose vectors for each of the plurality of atomic actions.

18. The system in accordance with any of Claims 12 to 17 wherein the features further comprise a spatial feature, and wherein the feature extractor derives each of a set of shape, motion or pose vectors for each of two or more regions of a bounding box within each of the plurality of atomic actions, the bounding box in each of the plurality of atomic actions configured to include all of a subject pictured in the one of the plurality of atomic actions, and wherein the motion recognition module convolves histograms of each of the shape, motion or pose descriptors to generate a resultant histogram.

19. The system in accordance with any of Claims 12 to 18 wherein the motion recognition module normalizes the histograms of each of the plurality of atomic actions to sum up to unity.

20. The system in accordance with any of Claims 11 to 19 wherein the motion recognition module further comprises a bag-of-words model module for K-means clustering of all of the atomic actions to generate a distance weighted bag-of-automatic-actions model of the video sequence.

1/9

100

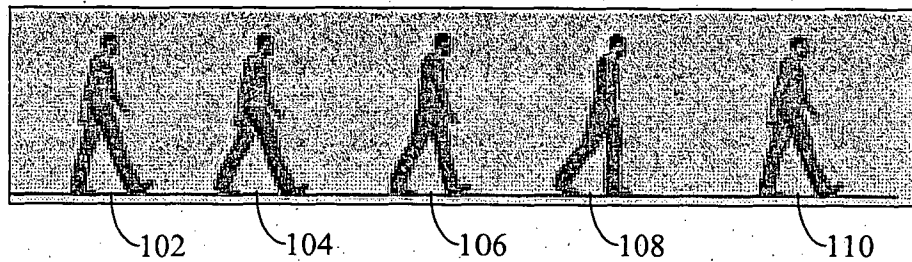


FIG. 1A

120

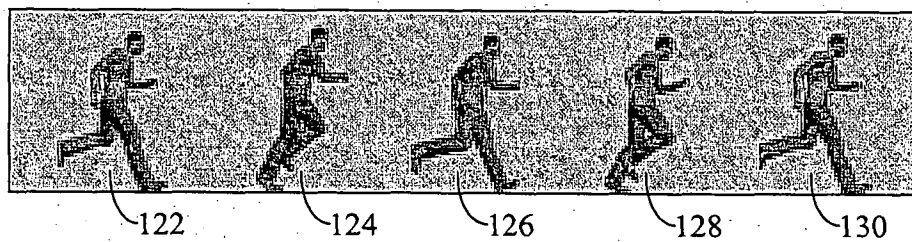


FIG. 1B



140

FIG. 1C



150

FIG. 1D

2/9



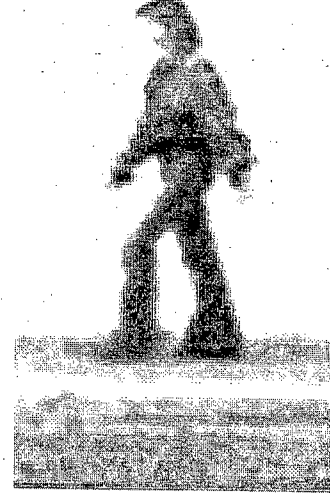
200

FIG. 2A



210

FIG. 2B



220

FIG. 2C

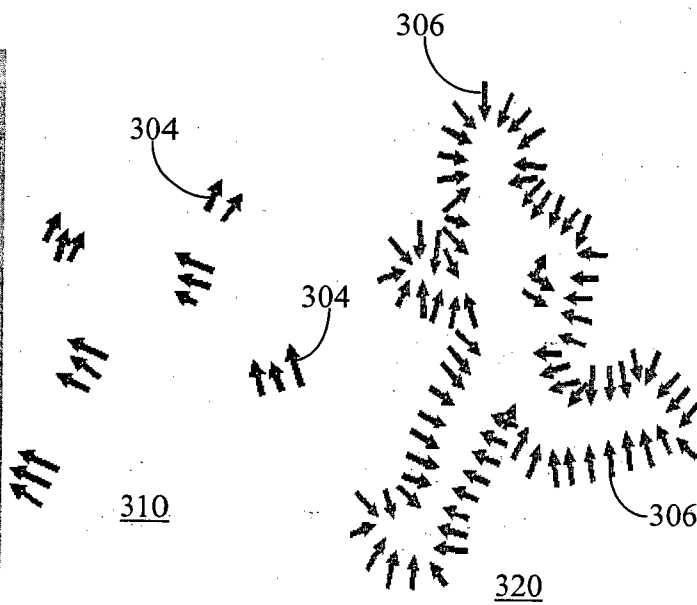


300

304

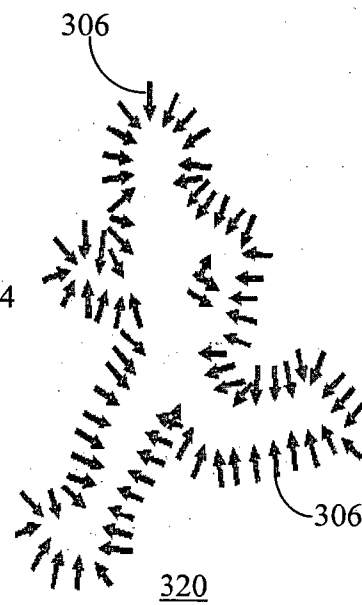
306

FIG. 3A



310

FIG. 3B



320

FIG. 3C

3/9

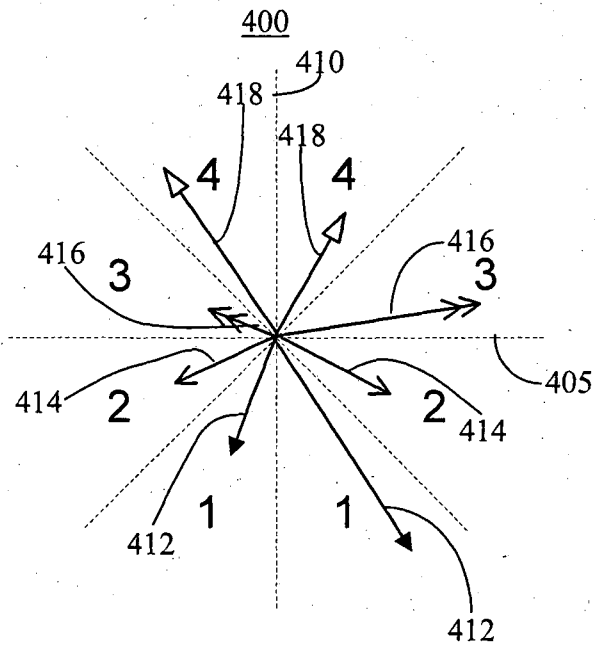


FIG. 4A

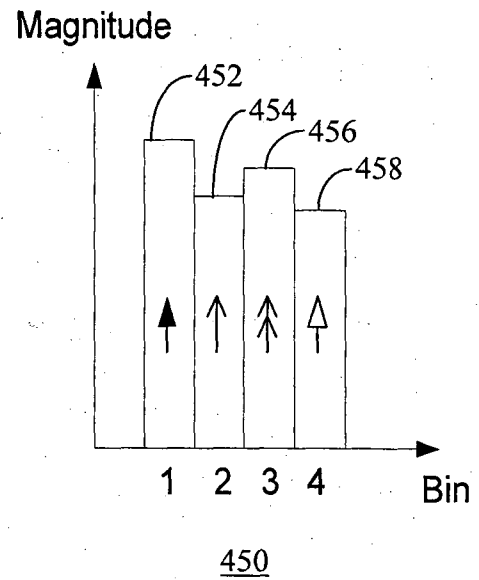


FIG. 4B

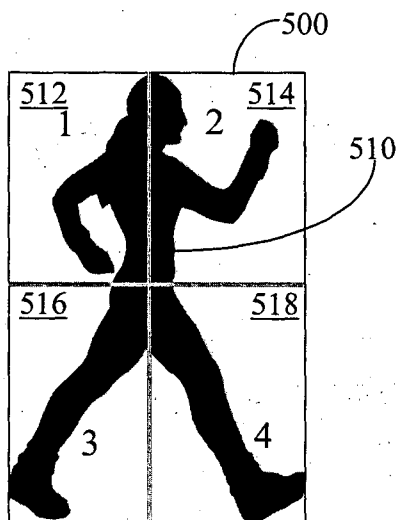


FIG. 5A

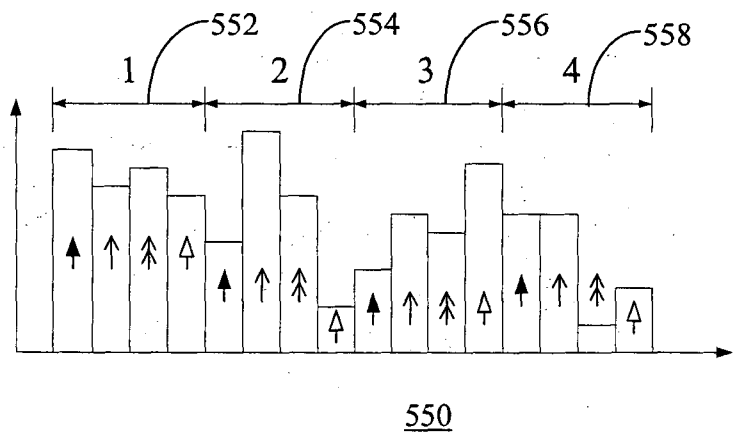
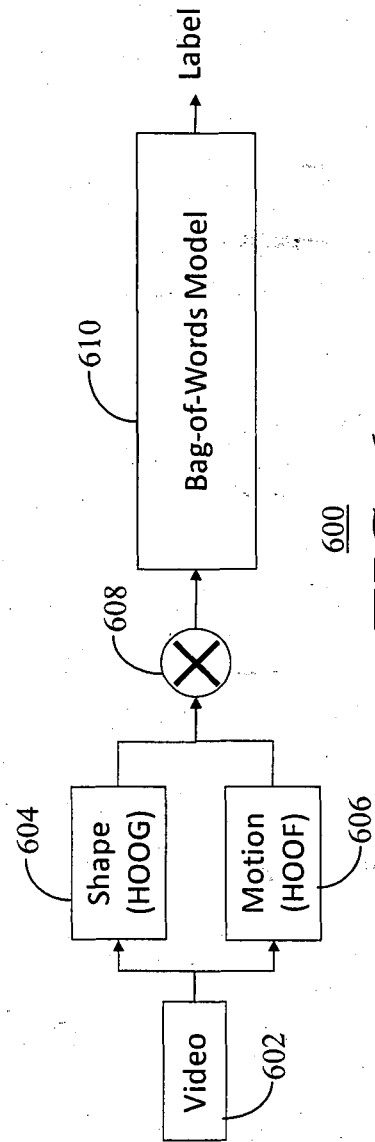
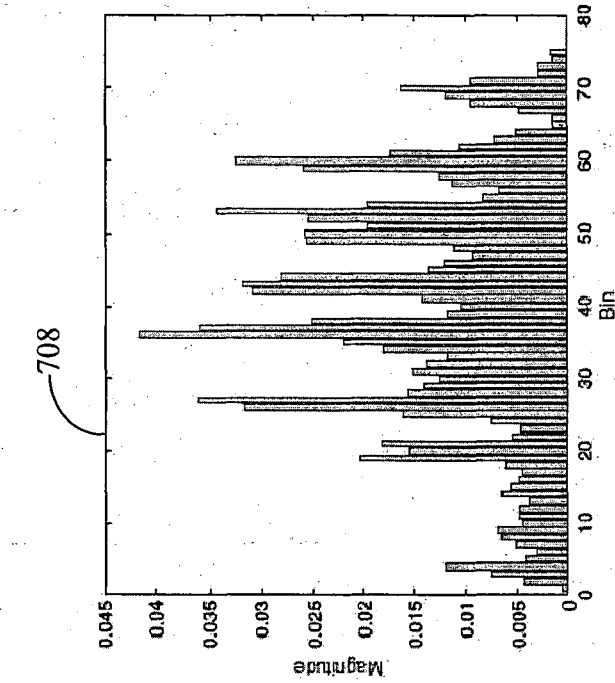


FIG. 5B



600

FIG. 6



700

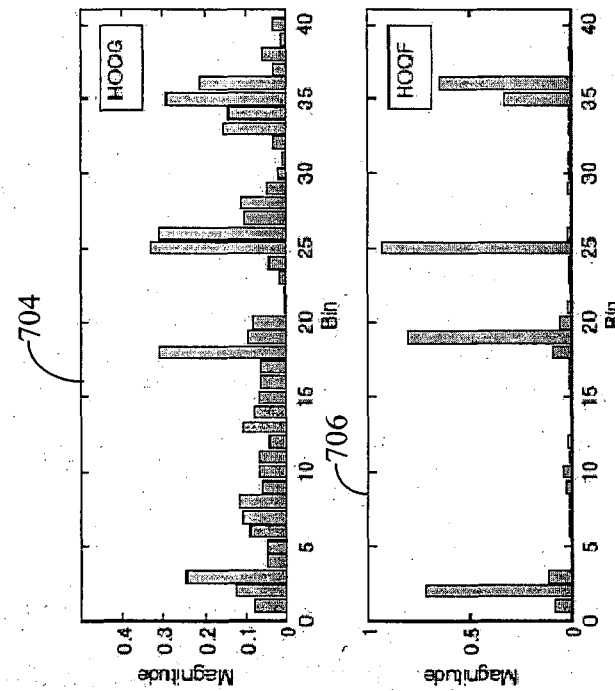
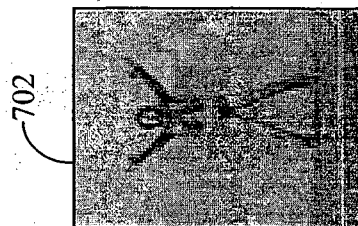


FIG. 7



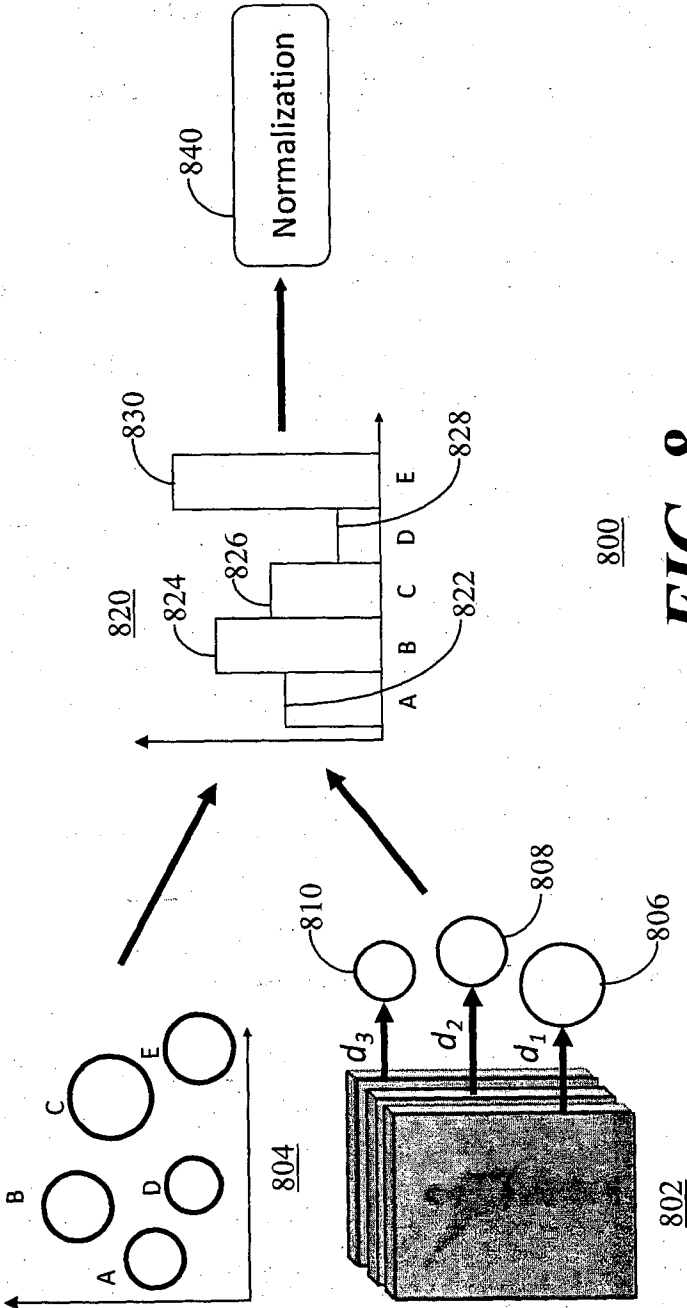


FIG. 8

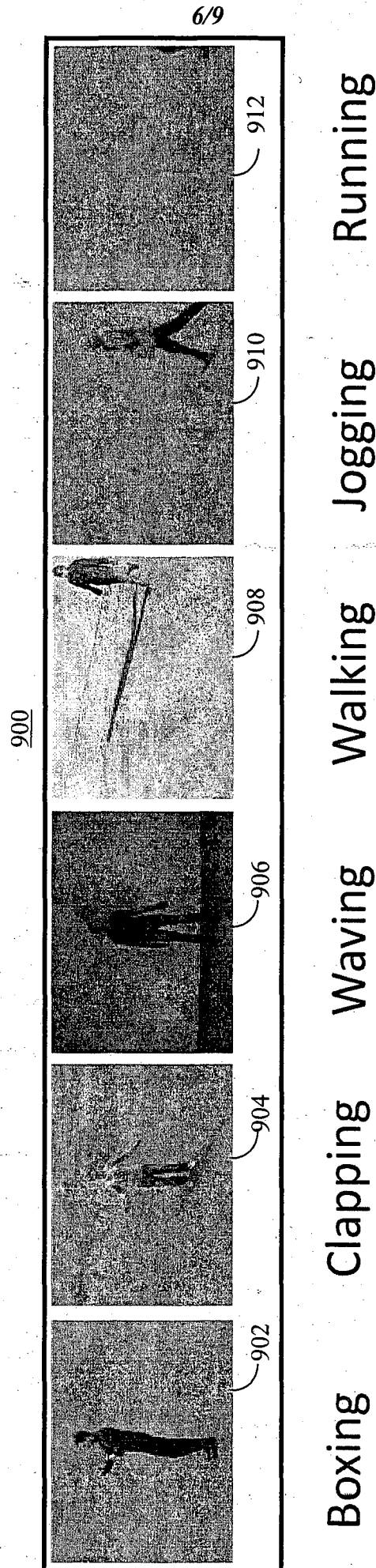


FIG. 9A

7/9

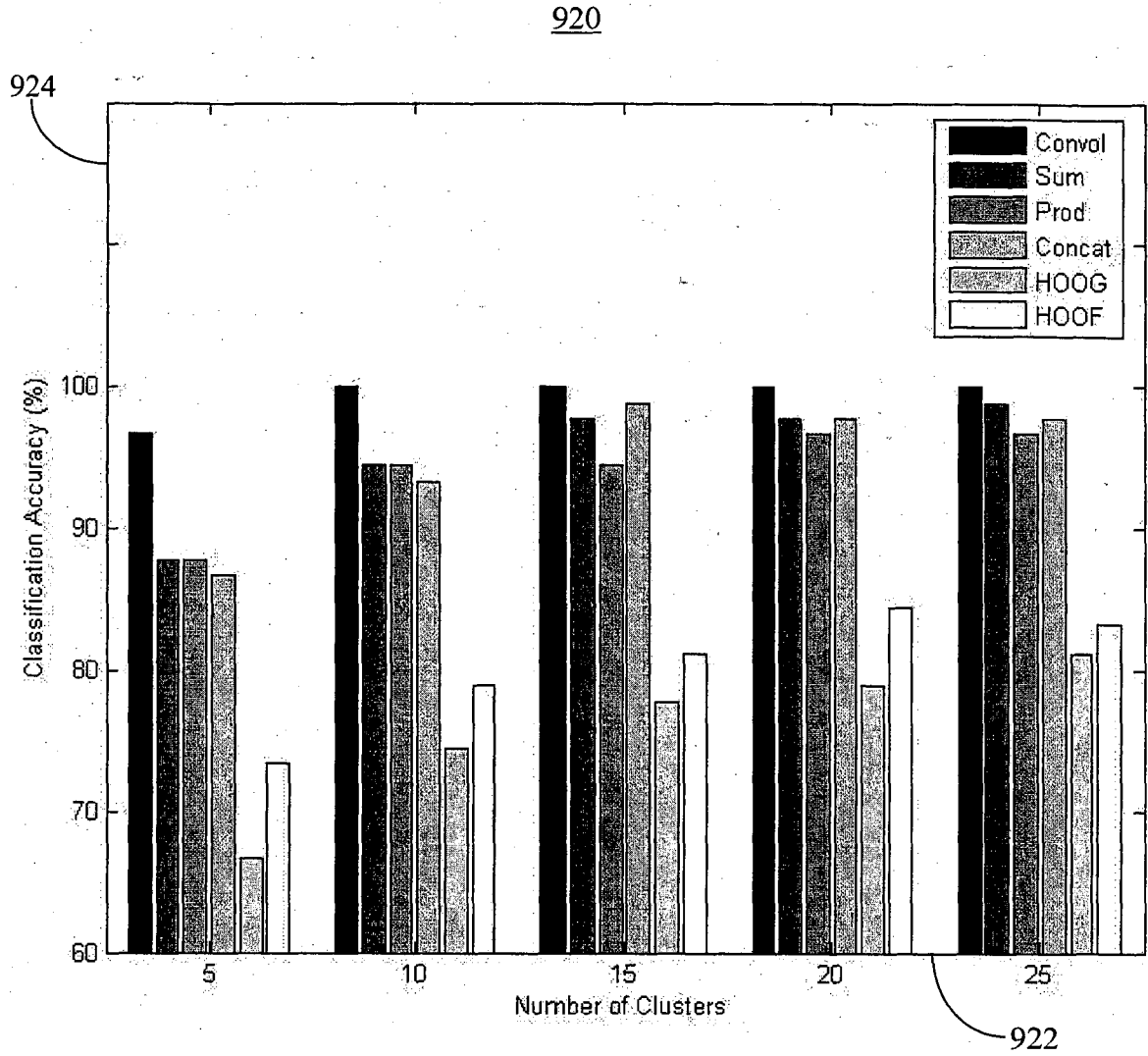
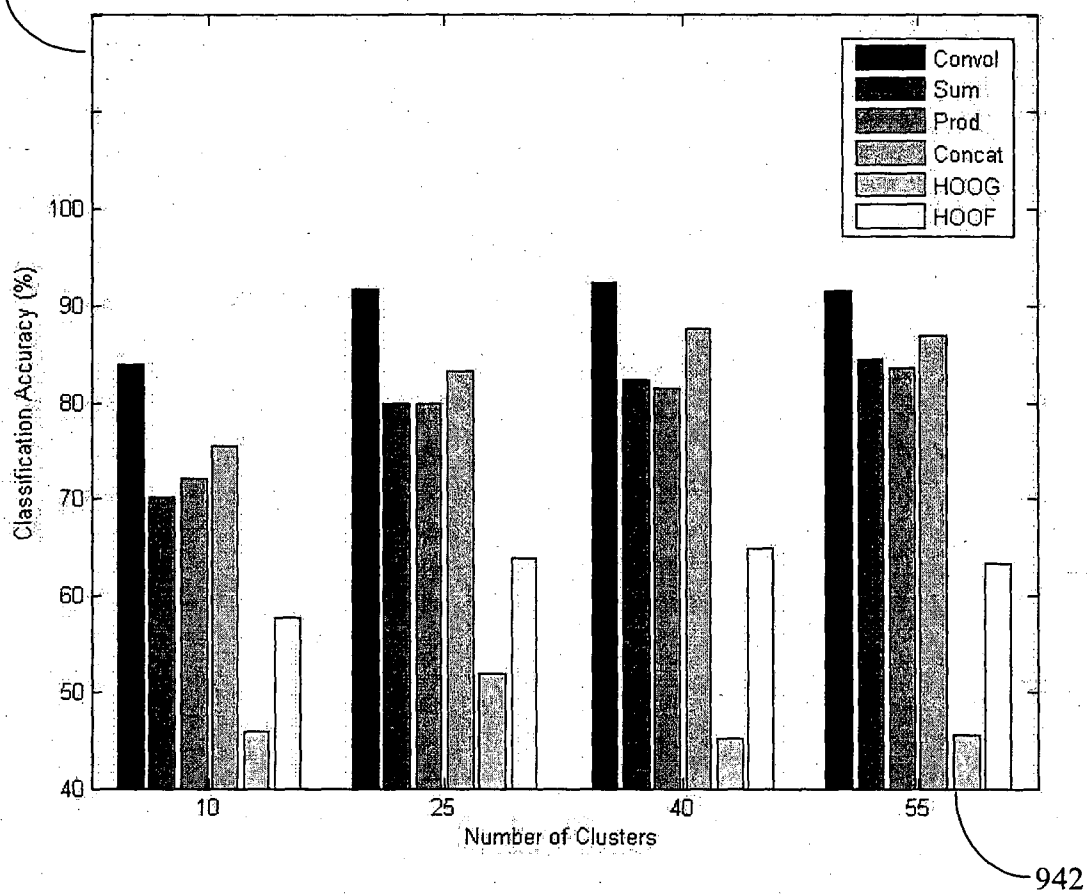


FIG. 9B

8/9

940

944

**FIG. 9C**

9/9

1000

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing 1002	83.84	13.13	2.02	1.01		
Handclapping	5.05	93.94	1.01			
Handwaving	3.00	10.00	86.00	1.00		
Jogging				94.00	6.00	
Running				1.00	99.00	
Walking				2.00		98.00

1004

FIG. 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2014/000275

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/46 (2006.01)

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

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

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

WPI, EPODOC & search terms: G06K, sequence, frame, segment, window, slide, image, video, surveil, vision, human, body, limb, pose, posture, motion, flow, shape, outline, contour, boundary, feature, character, vector, descriptor, action, behaviour, behavior, decompose, divide, separate, extract, recogni[s,z]e, match, identify, classify, categori[s,z]e, element, atomic, single, individual, isolate, convolution, convolve, combine, reverb, normali[s,z]e and similar terms; **The Lens** & search terms: "human motion", recognition, feature, extraction, shape, vector, descriptor, atomic; **Google, Google Patents, Google Scholar** & search terms: "human motion", recognition, feature, extraction, shape, vector, descriptor, atomic, (convolution OR convolve), histogram; **USPTO Patent Collection and Application** & search terms: "human motion", recognition, feature, extraction, shape, vector, descriptor, atomic, (convolution OR convolve), histogram; **Google**: Applicant/Inventor name search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Documents are listed in the continuation of Box C	

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

* "A"	Special categories of cited documents: document defining the general state of the art which is not considered to be of particular relevance	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E"	earlier application or patent but published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means	"&"	document member of the same patent family
"P"	document published prior to the international filing date but later than the priority date claimed		
Date of the actual completion of the international search 2 September 2014		Date of mailing of the international search report 02 September 2014	
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA Email address: pct@ipaustalia.gov.au		Authorised officer Benjamin Lam AUSTRALIAN PATENT OFFICE (ISO 9001 Quality Certified Service) Telephone No. +61 2 6225 6121	

INTERNATIONAL SEARCH REPORT		International application No.
C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		PCT/SG2014/000275
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	IKIZLER, N., 'Understanding Human Motion: Recognition and Retrieval of Human Activities, A dissertation submitted to the Department of Computer Engineering and the Institute of Engineering and Science of Bilkent University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy', May 2008 [retrieved on 28 August 2014] Retrieved from internet <URL: https://web.archive.org/web/20120311011636/http://www.es.bilkent.edu.tr/~duvgulu/Thesis/NazliIkizlerThesis.pdf> published on 11 March 2012 as per Wayback Machine Entire document (see particularly figures 1.1-6.12; page iv, Abstract; page 6, lines 5-7; page 12, section 2.2; pages 13-17, sections 2.2.2, 2.2.3, 2.2.4, 2.2.7; pages 20-42, entire Chapter 3, sections 3.1-3.3.3; page 59, paragraph starting with "In Fig. 4.9"; page 106, lines 9-12)	1-20
A	TURAGA, P., 'Machine Recognition of Human Activities: A Survey', Circuits and Systems for Video Technology, IEEE Transactions on Circuits and Systems for Video Technology (Volume: 18, Issue: 11), November 2008, pages 1473-1488	
A	NAYAK, N. et al., 'Motion pattern analysis for modeling and recognition of complex human activities. Guide to Video Analysis of Humans: Looking at People', New York, Springer-Verlag 2011, pages 289-310	
A	THURAU, C., 'Behavior Histograms for Action Recognition and Human Detection', Human Motion -- Understanding, Modeling, Capture and Animation, Second Workshop, Human Motion 2007, Rio de Janeiro, Brazil, 20 October 2007, pages 299-312	
A	US 2010/0310157 A1 (KIM et al.) 09 December 2010	
A	US 2012/0070070 A1 (LITVAK) 22 March 2012	
Form PCT/ISA/210 (fifth sheet) (July 2009)		

INTERNATIONAL SEARCH REPORT Information on patent family members		International application No. PCT/SG2014/000275	
This Annex lists known patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.			
Patent Document/s Cited in Search Report		Patent Family Member/s	
Publication Number	Publication Date	Publication Number	Publication Date
US 2010/0310157 A1	09 December 2010	US 8755593 B2	17 Jun 2014
		KR 20100131060 A	15 Dec 2010
US 2012/0070070 A1	22 March 2012	US 8582867 B2	12 Nov 2013
		US 2014037191 A1	06 Feb 2014
End of Annex			
Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001. Form PCT/ISA/210 (Family Annex)(July 2009)			