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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 January 2020 (09.01.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/008667 A1

(51) International Patent Classification:
G06K 9/00 (2006.01)

(21) International Application Number:
PCT/JP2019/002608

(22) International Filing Date:
22 January 2019 (22.01.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16/028,452 06 July 2018 (06.07.2018) US

(71) Applicant: **MITSUBISHI ELECTRIC CORPORATION** [JP/JP]; 7-3, Marunouchi 2-chome, Chiyoda-ku, Tokyo, 1008310 (JP).

(72) Inventor: **JONES, Michael**; c/o Mitsubishi Electric Research Laboratories, Inc., 201 Broadway, Cambridge, Massachusetts 02139 (US).

(74) Agent: **SOGA, Michiharu** et al.; S.SOGA & CO., 8th Floor, Kokusai Building, 1-1, Marunouchi 3-chome, Chiyoda-ku, Tokyo 1000005 (JP).

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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,

(54) Title: SYSTEM AND METHOD FOR VIDEO ANOMALY DETECTION

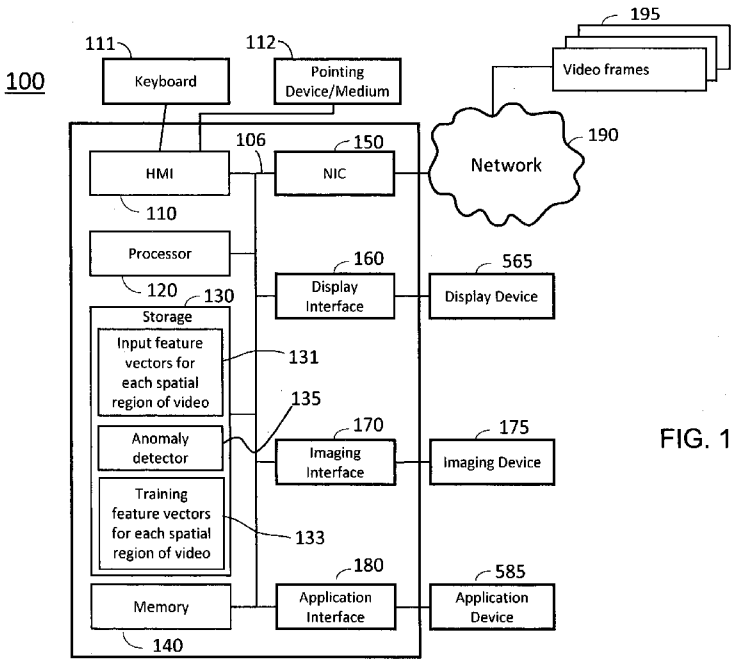

FIG. 1

(57) **Abstract:** A system for video anomaly detection partitions the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions, and determines blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images. Next, the system extracts a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors, computes a smallest distance between each input feature vector and the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances, and compares each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

WO 2020/008667 A1

[Continued on next page]

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).

Published:

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

[DESCRIPTION]

[Title of Invention]

SYSTEM AND METHOD FOR VIDEO ANOMALY DETECTION

[Technical Field]

[0001]

This invention relates generally to computer vision and more particularly to detecting motion anomalies in video.

[Background Art]

[0002]

Closed circuit television (CCTV) is widely used for security, transport and other purposes. Example applications include the observation of crime or vandalism in public open spaces or buildings (such as hospitals and schools), intrusion into prohibited areas, monitoring the free flow of road traffic, detection of traffic incidents and queues, detection of vehicles travelling the wrong way on one-way roads.

[0003]

The monitoring of CCTV displays (by human operators) is a very laborious task however and there is considerable risk that events of interest may go unnoticed. This is especially true when operators are required to monitor a number of CCTV camera outputs simultaneously. As a result, in many CCTV installations, video data is recorded and only inspected in detail if an event is known to have taken place. Even in these cases, the volume of recorded data may be large and the manual inspection of the data may be laborious. Consequently, there is a need for automatic devices to process video images to detect when there is an event of interest. Such a detection is referred herein as a video anomaly detection and can be used to draw the event to the immediate attention of an operator, to place an index mark in recorded video and/or to trigger selective recording of CCTV data.

[0004]

The problem of video anomaly detection is to automatically detect activity in part of a video that is very different from activities seen in normal (training) video of the same scene. For example, the video may be of a street scene with people walking along a sidewalk. Anomalous activity to be detected might be people fighting or climbing over a fence.

[0005]

There have been various approaches to the video anomaly detection problem published in the computer vision literature. One class of approaches typified by the paper, "Learning Temporal Regularity in Video Sequences" by Hasan et al. (CVPR 2016) uses a convolutional neural network autoencoder to learn the typical appearances and motions that occur in the training video. The autoencoder learns to reconstruct typical windows of the training video. To detect anomalies, the autoencoder is used to reconstruct windows of the testing video. Frames with high reconstruction error are flagged as anomalous. The main drawback of this method is that rare but normal activity that occurs in the training video is not well modeled which results in lots of false positive anomaly detections in testing videos.

[0006]

Another class of approaches is typified by the paper, "Abnormal Event Detection at 150 FPS in MATLAB", by C. Lu et al. (ICCV 2013) is based on reconstructing a feature vector computed from part of an input video from feature vectors of the training video. However, this class of approaches is error-prone and computationally expensive and also can miss rare but normal activity that occurs in the training video.

[0007]

A third class of approaches to video anomaly detection models the probability distribution of features of the video. "Anomaly Detection in Extremely

Crowded Scenes Using Spatio-Temporal Motion Pattern Models" by Kratz and Nishino (CVPR 2009) is an example of this approach. However, this approach can also miss the rare but normal activity that occurs in the training video.

[0008]

Accordingly, there is still a need for a system and a method for detecting motion anomalies in the input video.

[Summary of Invention]

[0009]

A typical scene that is of interest for surveillance can include virtually unlimited number of motion variations that is considered normal for that scene. To that end, it can be impractical to compare the acquired video of the scene with all possible variation of the normal motion of the scene to detect anomalies in the video. A number of anomaly detection methods consider this not as a problem but as a fact of life. To that end, various anomaly detection methods aim to come up with an aggregation of the normal motion in the scene and use that aggregation in the anomaly detection. The aggregation can have different forms, such as parameters of a neural network, motion reconstruction techniques or probability distributions. However, the computation of the motion aggregation can be expensive, the motion aggregation can fail to recognize rare but normal motion in the scene, or both.

[0010]

It is an object of some embodiments to use direct comparison between the motion in the input video and the motion in the training video for anomaly detection. Such a direct comparison, e.g., the comparison based on Euclidean distance, is computationally efficient and can consider even rare normal motions. For example, if usual normal motion on the street is walking, the motion aggregation based method can distinguish the walking from falling. However, if

the normal motion includes a motion of a person tying her shoelaces, the aggregation-based methods would consider that motion as an anomaly even if the motion of tying shoes is occasionally present in the training video. In contrast, the direct distance computation methods can recognize this tying shoes motion as normal, if such type of motion is present in the training video.

[0011]

Some embodiments are based on the recognition that the aggregation techniques of video anomaly detection are post-processing techniques, i.e., the techniques that are used after the motions in the training video are computed. Some embodiments are based on the realization that those post-processing techniques can be replaced with pre-processing techniques used to determine the motions in the first place. Specifically, some embodiments are based on realization that the aggregation problem in video anomaly detection can be addressed when instead of providing an aggregation of the motions computed from the training video, the computation of the motions themselves is performed to reduce a search space for the computed motions while preserving the accuracy and advantages of the direct comparison.

[0012]

Specifically, some embodiments are based on the realization that such a search-space efficient motion computation can be accomplished via the blurred, thresholded difference of two consecutive frames in the training and the input video. For example, the thresholded difference image of two consecutive video frames capture the motion in those frames. Any pixel that changes significantly (i.e. for which the absolute value of the difference is above a threshold) is assigned a value one in the thresholded difference image and a value zero otherwise. In the blurred, thresholded difference image, this motion image is blurred, e.g., by convolving the thresholded difference image with a Gaussian kernel. Some

embodiments are based on realization that if two thresholded difference images capture similar but different motions, then this difference results in two different thresholded difference images. However, when two blurred, thresholded difference images are determined for the same similar but different motions, the blurring can smooth out the difference such that these two blurred thresholded difference images can be considered similar.

[0013]

This realization allows for a compact and efficient method of storing all of the normal motions that are present in the training video of the scene. The training video is divided into different spatio-temporal regions. For example, in some implementations, a spatio-temporal region of the video is represented by a fixed-length sequence of T blurred, thresholded difference images within the spatial region computed over T+1 consecutive frames of the training video. The sequence of T blurred, thresholded difference images within a spatial region are stacked together and vectorized to yield a feature vector.

[0014]

For a training video, a number of feature vectors can be very similar due partly to the blurring operation. Two feature vectors are considered duplicates of each other if the distance between them is less than a threshold, referred herein as a duplication threshold. In such a manner, when only unique feature vectors in each spatial region of the training video are stored, the search space of the motion is reduced making that search space practical for direct motion comparison.

[0015]

Accordingly, one embodiment discloses a system for video anomaly detection that includes a memory to store a set of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene and a processor to accept an input video of a scene

through an input interface and to detect an anomaly in the input video of the scene based on comparison of the input video with the training video, i.e., comparison with the training feature vectors stored in the memory.

[0016]

Specifically, the processor is configured to partition the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions; determine sequences of blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of input feature vectors; compute a distance between each input feature vector and each of the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and compare each distance from the set of distances with an anomaly detection threshold to detect an anomaly in the input video of the scene.

[0017]

For example, the processor can detect the anomaly in the input video of the scene when at least one distance in the set of distances is greater than the anomaly detection threshold. In some implementations, the processor executes a control action in response to detecting the anomaly. For example, the control action can draw the event to the immediate attention of an operator, place an index mark in recorded video, trigger selective recording of CCTV data, and/or cause an execution of a security application.

[0018]

Some embodiments are based on recognition that when the blurred, thresholded difference images are determined to reduce a search space of anomaly

detection, a special need arise for selecting parameters of blurring and/or for identifying duplicates of the feature vectors. For example, some embodiments are based on realization that there is a need to determine a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region. Such a minimum distance is referred herein as a distance threshold.

[0019]

The distance threshold allows reducing a number of training feature vectors stored in the memory, while preserving accuracy of anomaly detection. Some embodiments are based on realization that the values of the distance threshold can depend on the value of the anomaly detection threshold used for direct comparison between the input and training feature vector. To that end, in some embodiments the distance threshold is a function of the anomaly detection threshold.

[0020]

For example, in some embodiments the distance threshold is proportional to the anomaly detection threshold according to a coefficient of proportionality. For example, in some implementations, the distance threshold is equal to or greater than the anomaly detection threshold, i.e., the coefficient of proportionality is equal to or greater than one. This embodiment can advantageously balance the accuracy of the anomaly detection with its speed, because larger values of the coefficient of proportionality reduce the memory requirement to store the training feature vectors, increase the speed of comparison, but decrease the accuracy of anomaly detection. In some embodiments, the coefficient of proportionality is between one and two inclusively to reduce the negative effect of the search space reduction on the accuracy of the anomaly detection.

[0021]

Additionally, or alternatively, in one embodiment, the distance threshold is greater than a median distance of the training feature vectors stored in the memory

and corresponding to the same spatial region. For example, some implementations calculate a mean of distances between each training feature vector and all other training feature vectors in the same spatial region and increase the mean by a standard deviation to produce the distance threshold. For example, one embodiment calculates the mean distance, m, and the standard deviation, s, and determines the distance threshold T_2 as $T_2 = m + c*s$ where c is a positive constant, e.g., one, two or three. These embodiments take advantage of using the mean and standard deviation, which is better justified by statistics.

[0022]

Additionally, or alternatively, in some embodiments, the value of the blur kernel is inversely proportional to the distance threshold. In other words, the more blur is used the smaller the distance threshold is, because more blur makes video patches more similar. For example, in some implementations, the distance threshold T_1 is inversely proportional to the blur kernel according to $T1 = c/n$, where c is a constant, i.e., the coefficient of the inverse proportionality, and n is the length in pixels of the square blur kernel.

[0023]

Accordingly, one embodiment discloses a system for video anomaly detection, including an input interface to accept an input video of a scene; a memory to store sets of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene; and a processor configured to partition the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions; determine blurred, thresholded difference images for each of the

input spatio-temporal regions to produce a set of blurred, thresholded difference images; extract a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors; compute a smallest distance between each input feature vector and the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and compare each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

[0024]

Another embodiment discloses a method for video anomaly detection, wherein the method uses a processor coupled to a memory storing sets of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene, wherein the processor is coupled with stored instructions implementing the method, wherein the instructions, when executed by the processor carry out at least some steps of the method. The method includes accepting an input video of a scene; partitioning the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions; determining blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images; extracting a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors; computing a smallest distance between each input feature vector and the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and comparing each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

[0025]

Yet another embodiment discloses a non-transitory computer readable storage medium embodied thereon a program executable by a processor for performing a method, the method includes accepting an input video of a scene; partitioning the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions; determining blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images; extracting a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors; computing a smallest distance between each input feature vector and training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and comparing each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

[Brief Description of the Drawings]

[0026]

[Fig. 1]

Figure 1 shows a block diagram of an image processing system for detecting anomalies in videos in accordance with some embodiments.

[Fig. 2A]

Figure 2A shows a flowchart of a method for detecting anomalies in frames of a video according to some embodiments.

[Fig. 2B]

Figure 2B shows a flow chart of a method of preparing sets of training

feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene according to one embodiment.

[Fig. 3]

Figure 3 shows a schematic of determining a blurred, thresholded difference image used by some embodiments to reduce the search space.

[Fig. 4]

Figure 4 shows a schematic of creating a feature vector from a set of blurred, thresholded difference images computed from a spatio-temporal region of video according to one embodiment.

[Fig. 5]

Figure 5 shows a schematic of a nearest neighbor search method used by some embodiments to find the closest training feature vector to a testing feature vector representing a sequence of blurred, thresholded difference images.

[Description of Embodiments]

[0027]

Figure 1 shows a block diagram of an image processing system 100 for detecting anomalies in videos in accordance with some embodiments. The image processing system 100 includes a processor 120 configured to execute stored instructions, as well as a memory 140 that stores instructions that are executable by the processor. The processor 120 can be a single core processor, a multi-core processor, a computing cluster, or any number of other configurations. The memory 140 can include random access memory (RAM), read only memory (ROM), flash memory, or any other suitable memory systems. The processor 120 is connected through a bus 106 to one or more input and output devices. These instructions implement a method for detecting anomalies in a video sequence.

[0028]

In various embodiments, the anomaly detection produces a set of bounding

boxes indicating the locations and sizes of any anomalies in each video frame. The image processing system 100 is configured to detect anomalies in a video using an anomaly detector 135 that compares feature vectors 131 of spatio-temporal regions of the input video of a scene to feature vectors 133 from spatio-temporal regions of the training video of the same scene. For example, a storage device 130 can be adapted to store the sets of feature vectors computed from different spatial regions of training and/or input video frames. The storage device 130 can be implemented using a hard drive, an optical drive, a thumb drive, an array of drives, or any combinations thereof.

[0029]

Features, as used herein, are information or values extracted from the video or frames of the video. In some implementations, the features are the function of intensities of the frames of the video. A feature vector is a data structure storing the features, e.g., intensities. The feature vector can be one or multi-dimensional. For example, in some embodiments, the feature vector stores intensities of blurred, thresholded difference images produced from the input and the training videos.

[0030]

Various embodiments can use different spatio-temporal partitions of the video of the scene. However, in various implementations, the spatio-temporal partitions of the input video are identical to the spatio-temporal partitions of the training video to streamline the comparison.

[0031]

In some implementations, a human machine interface 110 within the image processing system 100 connects the system to a keyboard 111 and pointing device 112, wherein the pointing device 112 can include a mouse, trackball, touchpad, joy stick, pointing stick, stylus, or touchscreen, among others. The image processing system 100 can be linked through the bus 106 to a display interface 160 adapted to

connect the image processing system 100 to a display device 565, wherein the display device 565 can include a computer monitor, camera, television, projector, or mobile device, among others.

[0032]

The image processing system 100 can also be connected to an imaging interface 170 adapted to connect the system to an imaging device 175. In one embodiment, the frames of input video on which the anomaly detector is run are received from the imaging device. The imaging device 175 can include a video camera, computer, mobile device, webcam, or any combination thereof.

[0033]

A network interface controller 150 is adapted to connect the image processing system 100 through the bus 106 to a network 190. Through the network 190, the video frames 195, e.g., frames of the training and/or video can be downloaded and stored within the computer's storage system 130 for storage and/or further processing.

[0034]

In some embodiments, the image processing system 100 is connected to an application interface 180 through the bus 106 adapted to connect the image processing system 100 to an application device 585 that can operate based on results of anomaly detection. For example, the device 585 is a surveillance system that uses the locations of detected anomalies to alert a security guard to investigate further.

[0035]

Figure 2A shows a flowchart of a method for detecting anomalies in frames of a video according to some embodiments. In step S1, the input video 215 is partitioned into spatio-temporal regions 210. Each spatio-temporal region can be defined by a bounding box in a video frame defining the spatial extent and a fixed

number of frames, defining the temporal extent. Different spatio-temporal regions may be overlapping. The union of all spatio-temporal regions cover the entire video sequence. In step S2, blurred, thresholded difference images 220 are computed for each spatio-temporal region. In step S3, feature vectors 230 are created from the blurred, thresholded difference images. For example, one embodiment simply concatenates all pixels values into a single vector. In step S4, the distance of each feature vector is computed to a set of training feature vectors computed for the same spatial region from training video of normal activity from the same camera as the testing video to find the minimum distance 240, i.e., the distance to its nearest neighbor. In step S5, the minimum distance is assigned as the anomaly score 250 for this spatio-temporal region of the input video.

[0036]

In step S6, each anomaly score is compared to a threshold and if the score is above a threshold then the spatio-temporal region is classified as anomalous. Otherwise the region is classified as normal. For example, one embodiment is configured to detect the anomaly in the input video of the scene when at least one distance in the set of distances is greater than the anomaly detection threshold; and execute a control action in response to detecting the anomaly. The steps of the method are performed for each of the different spatio-temporal region 210.

[0037]

Figure 2B shows a flow chart of a method of preparing sets of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene according to one embodiment. The training feature vectors are considered to represent a normal behavior in the scene. In other words, if the input frame is represented by a frame in the training video, the scene in the input frame is considered normal. Otherwise, that scene is considered anomalous.

[0038]

The training features vectors are determined in a manner similar to determining the input feature vectors. For example, the embodiment is configured to partition S11 the training video 265 into the spatio-temporal regions 270 of the training video. The dimensions of the resolution of the spatio-temporal regions 210 corresponds to the dimensions and/or resolutions of the spatio-temporal regions 270. Next, the embodiment determines S22 blurred difference images for each of the training spatio-temporal regions 270 to produce a set of training blurred difference images 275 and extracts S33 features from each blurred difference image in the set of training blurred difference images to produce training feature vectors 280.

[0039]

The embodiment is configured to compute S44 a distance between each training feature vector corresponding to the same spatial region in the scene to produce a set of distances 285 between training feature vectors. Next, the embodiment prunes similar training feature vector to reduce the computational burden of the comparison. Specifically, the embodiment selects S55 a training feature vector in the set training feature vectors when all distances between the selected training feature vector and corresponding feature vectors in the set of training feature vectors is above a distance threshold defining a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region.

[0040]

In some embodiments, to save the memory space and improve the comparisons, the minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region is a function of the anomaly detection threshold. In such a manner, those thresholds are jointly used to

improve the anomaly detection. For example, in one implementation, a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region is equal to or greater than the anomaly detection threshold.

[0041]

Additionally, or alternatively, some embodiments consider only information from the training video itself to determine a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region. For example, in one implementations the minimum distance is greater than a median distance of the training feature vectors stored in the memory and corresponding to the same spatial region. This embodiment provides data-driven approach for determining the minimum distance. In addition, in some implementations the anomaly detection threshold is determined as a function of the minimum distance determined by the data-driven approach. In such a manner, both minimum distance and the anomaly detection threshold are automatically adapted for different scenes.

[0042]

In such a manner, some embodiments to use direct comparison between the motion in the input video and the motion in the training video for anomaly detection. Such a direct comparison, e.g., the comparison based on Euclidean distance, is computationally efficient and can consider even rare normal motions. For example, if usual normal motion on the street is walking, the motion aggregation based method can distinguish the walking from falling. However, if the normal motion includes a motion of a person tying her shoelaces, the aggregation-based methods would consider that motion as an anomaly even if the motion of tying shoes is occasionally present in the training video. In contrast, the direct distance computation methods can recognize this tying shoes motion as

normal, if such type of motion is present in the training video.

[0043]

Some embodiments are based on the recognition that the aggregation techniques of video anomaly detection are post-processing techniques, i.e., the techniques that are used after the motions in the training video are computed. Some embodiments are based on the realization that those post-processing techniques can be replaced with pre-processing techniques used to determine the motions in the first place. Specifically, some embodiments are based on realization that the aggregation problem in video anomaly detection can be addressed when instead of providing an aggregation of the motions computed from the training video, the computation of the motions themselves is performed to reduce a search space for the computed motions while preserving the accuracy and advantages of the direct comparison.

[0044]

Specifically, some embodiments are based on the realization that such a search-space efficient motion computation can be accomplished via the blurred, thresholded difference of two consecutive frames in the training and the input video. For example, the thresholded difference image of two consecutive video frames capture the motion in those frames. Any pixel that changes significantly (i.e. for which the absolute value of the difference is above a threshold) is assigned a value one in the thresholded difference image and a value zero otherwise. In the blurred, thresholded difference image, this motion image is blurred, e.g., by convolving the thresholded difference image with a Gaussian kernel. Some embodiments are based on realization that if two thresholded difference images capture similar but different motions, then this difference results in two different thresholded difference images. However, when two blurred, thresholded difference images are determined for the same similar but different motions, the blurring can

smooth out the difference such that these two blurred thresholded difference images can be considered similar.

[0045]

Figure 3 shows a schematic of determining a blurred, thresholded difference image used by some embodiments to reduce the search space. Given two video frames (A1 and A2), pixels in A1 are subtracted from pixels at the same location in A2 and the absolute value of each difference is computed. If the absolute value of the difference is greater than a threshold then the pixel of the thresholded difference image, A3, at that location is assigned the value 1, otherwise the pixel is assigned the value 0. Next, the thresholded difference image is blurred, for example using a Gaussian filter, to create a blurred, thresholded difference image (A4).

[0046]

Some embodiments are based on recognition that when the blurred, thresholded difference images are determined to reduce a search space of anomaly detection, a special need arises for selecting parameters of blurring and/or for identifying duplicates of the feature vectors. For example, some embodiments are based on realization that there is a need to determine a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region. Such a minimum distance is referred herein as a distance threshold.

[0047]

The distance threshold allows reducing a number of training feature vectors stored in the memory, while preserving accuracy of anomaly detection. Some embodiments are based on realization that the values of the distance threshold can depend on the value of the anomaly detection threshold used for direct comparison between the input and training feature vector. To that end, in some embodiments the distance threshold is a function of the anomaly detection threshold.

[0048]

For example, in some embodiments the distance threshold is proportional to the anomaly detection threshold according to a coefficient of proportionality. For example, in some implementations, the distance threshold is equal to or greater than the anomaly detection threshold, i.e., the coefficient of proportionality is equal to or greater than one. This embodiment can advantageously balance the accuracy of the anomaly detection with its speed, because larger values of the coefficient of proportionality reduce the memory requirement to store the training feature vectors, increase the speed of comparison, but decrease the accuracy of anomaly detection. In some embodiments, the coefficient of proportionality is between one and two inclusively to reduce the negative effect of the search space reduction on the accuracy of the anomaly detection.

[0049]

Additionally, or alternatively, in one embodiment, the distance threshold is greater than a median distance of the training feature vectors stored in the memory and corresponding to the same spatial region. For example, some implementations calculate a mean of distances between each training feature vector and all other training feature vectors in the same spatial region and increase the mean by a standard deviation to produce the distance threshold. For example, one embodiment calculates the mean distance, m, and the standard deviation, s, and determines the distance threshold T_2 as $T_2 = m + c*s$ where c is a positive constant, e.g., one, two or three. These embodiments take advantage of using the mean and standard deviation, which is better justified by statistics.

[0050]

Figure 4 shows a schematic of creating a feature vector from a set of blurred, thresholded difference images computed from a spatio-temporal region of video according to one embodiment. In this example, blurred, thresholded difference

images for 10 frames 410 of a spatio-temporal region are vectorized into a vector 420. For example, the value of each pixel in each blurred thresholded difference image is listed from top, left to bottom, right in each image and copied into a one-dimensional vector 420.

[0051]

In some embodiments, the value of the blur kernel is inversely proportional to the distance threshold. In other words, the more blur is used the smaller the distance threshold is, because more blur makes video patches more similar. For example, in some implementations, the distance threshold T_1 is inversely proportional to the blur kernel according to T1 = c/n, where c is a constant, i.e., the coefficient of the inverse proportionality, and n is the length in pixels of the square blur kernel.

[0052]

Figure 5 shows a schematic of a nearest neighbor search method used by some embodiments to find the closest training feature vector to a testing feature vector representing a sequence of blurred, thresholded difference images. In Figure 5, fv 510 is the input feature vector and each x_i 520 is a training feature vector. The nearest neighbor search 530 outputs the minimum distance, d, 540 between fv and the nearest x_i. Different embodiments use different nearest neighbor searches. For example, one embodiment uses brute force search to compare each input feature vector with each training feature vector. In some implementations, the nearest neighbor search 530 is an approximate nearest neighbor search, which is not guaranteed to find the minimum distance but may instead find a feature vector that is close to the minimum. Various nearest neighbor search algorithms know in the field could be used such as k-d trees, k-means trees, and locality sensitive hashing.

[0053]

The above-described embodiments of the present invention can be implemented in any of numerous ways. For example, the embodiments may be implemented using hardware, software or a combination thereof. When implemented in software, the software code can be executed on any suitable processor or collection of processors, whether provided in a single computer or distributed among multiple computers. Such processors may be implemented as integrated circuits, with one or more processors in an integrated circuit component. Though, a processor may be implemented using circuitry in any suitable format.

[0054]

Also, the embodiments of the invention may be embodied as a method, of which an example has been provided. The acts performed as part of the method may be ordered in any suitable way. Accordingly, embodiments may be constructed in which acts are performed in an order different than illustrated, which may include performing some acts simultaneously, even though shown as sequential acts in illustrative embodiments.

[0055]

Use of ordinal terms such as "first," "second," in the claims to modify a claim element does not by itself connote any priority, precedence, or order of one claim element over another or the temporal order in which acts of a method are performed, but are used merely as labels to distinguish one claim element having a certain name from another element having a same name (but for use of the ordinal term) to distinguish the claim elements.

[CLAIMS]

[Claim 1]

A system for video anomaly detection, comprising:

an input interface to accept an input video of a scene;

a memory to store sets of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene; and

a processor configured to

partition the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions;

determine blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images;

extract a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors;

compute the minimum distance between each input feature vector and the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and

compare each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

[Claim 2]

The system of claim 1, wherein the processor is configured to

detect the anomaly in the input video of the scene when at least one distance in the set of distances is greater than the anomaly detection threshold; and

execute a control action in response to detecting the anomaly.

[Claim 3]

The system of claim 1, wherein the training feature vectors stored in the memory for a spatial region are chosen so that the minimum distance among them is a function of the anomaly detection threshold.

[Claim 4]

The system of claim 1, wherein the training feature vectors stored in the memory for a spatial region are chosen so that the minimum distance among them is equal to or greater than the anomaly detection threshold.

[Claim 5]

The system of claim 1, wherein the training feature vectors stored in the memory for a spatial region are chosen so that the minimum distance among them is greater than a median distance between all possible pairs of training feature vectors for that spatial region.

[Claim 6]

The system of claim 1, wherein the processor to determine the training feature vectors is configured to

partition the training video into the spatio-temporal regions of the training video;

determine blurred, thresholded difference images for each of the training spatio-temporal regions to produce a set of training blurred, thresholded difference images;

extract features from each blurred, thresholded difference image in the set of training blurred, thresholded difference images to produce training feature vectors; and

compute a distance between each training feature vector corresponding to the same spatial region in the scene to produce a set of distances between training

feature vectors; and

select a training feature vector in the set of training feature vectors when all distances between the selected training feature vector and corresponding feature vectors in the set of training feature vectors are above a distance threshold defining a minimum distance among the training feature vectors stored in the memory and corresponding to the same spatial region.

[Claim 7]

The system of claim 6, wherein the processor, for determining the distance threshold, is configured to

calculate a mean of distances between all training feature vectors and the training feature vectors in the set of feature vectors; and

increase the mean by a standard deviation to produce the distance threshold.

[Claim 8]

The system of claim 1, wherein the processor, to determine a blurred, thresholded difference image, is configured to

determine a difference between subsequent video frames of the input video to produce a difference image;

threshold the difference image to produce a binary difference image; and

blur the binary difference image with a kernel to produce the blurred difference image.

[Claim 9]

The system of claim 8, wherein a type of the kernel is a Gaussian kernel.

[Claim 10]

The system of claim 8, wherein a value of the kernel is a function of a distance threshold defining a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region.

[Claim 11]

The system of claim 10, wherein the value of the kernel is inversely proportional to the distance threshold.

[Claim 12]

A method for video anomaly detection, wherein the method uses a processor coupled to a memory storing sets of training feature vectors extracted from blurred, thresholded difference images of spatio-temporal regions of a training video of the scene, wherein the processor is coupled with stored instructions implementing the method, wherein the instructions, when executed by the processor carry out at least some steps of the method, comprising:

accepting an input video of a scene;

partitioning the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video, wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions;

determining blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images;

extracting a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors;

computing a minimum distance between each input feature vector and the training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and

comparing each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

[Claim 13]

The method of claim 12, further comprising:

detecting the anomaly in the input video of the scene when at least one distance in the set of distances is greater than the anomaly detection threshold; and

executing a control action in response to detecting the anomaly.

[Claim 14]

The method of claim 12, further comprising:

partitioning the training video into the spatio-temporal regions of the training video;

determining blurred difference images for each of the training spatio-temporal regions to produce a set of training blurred difference images;

extracting features from each blurred difference image in the set of training blurred difference images to produce training feature vectors; and

computing a distance between each training feature vector corresponding to the same spatial region in the scene to produce a set of distances between training feature vectors;

selecting a training feature vector in the set of raining feature vectors when all distances between the selected training feature vector and corresponding feature vectors in the set of training feature vectors are above a distance threshold defining a minimum distance between the training feature vectors stored in the memory and corresponding to the same spatial region; and

storing the set of raining feature vectors in the memory.

[Claim 15]

A non-transitory computer readable storage medium embodied thereon a program executable by a processor for performing a method, the method comprising:

accepting an input video of a scene;

partitioning the input video into a set of input spatio-temporal regions according to parameters of the spatio-temporal regions of the training video,

wherein the parameters are indicative of a number of regions in each video frame defining a spatial dimension of each of the spatio-temporal regions and a number of video frames defining a temporal dimension of each of the spatio-temporal regions;

determining blurred, thresholded difference images for each of the input spatio-temporal regions to produce a set of blurred, thresholded difference images;

extracting a feature vector from each set of blurred, thresholded difference images to produce a set of input feature vectors; and

computing a smallest distance between each input feature vector and training feature vectors corresponding to the same spatial region in the scene to produce a set of distances; and

comparing each distance from the set of distances with an anomaly detection threshold to detect anomalies in the input video of the scene.

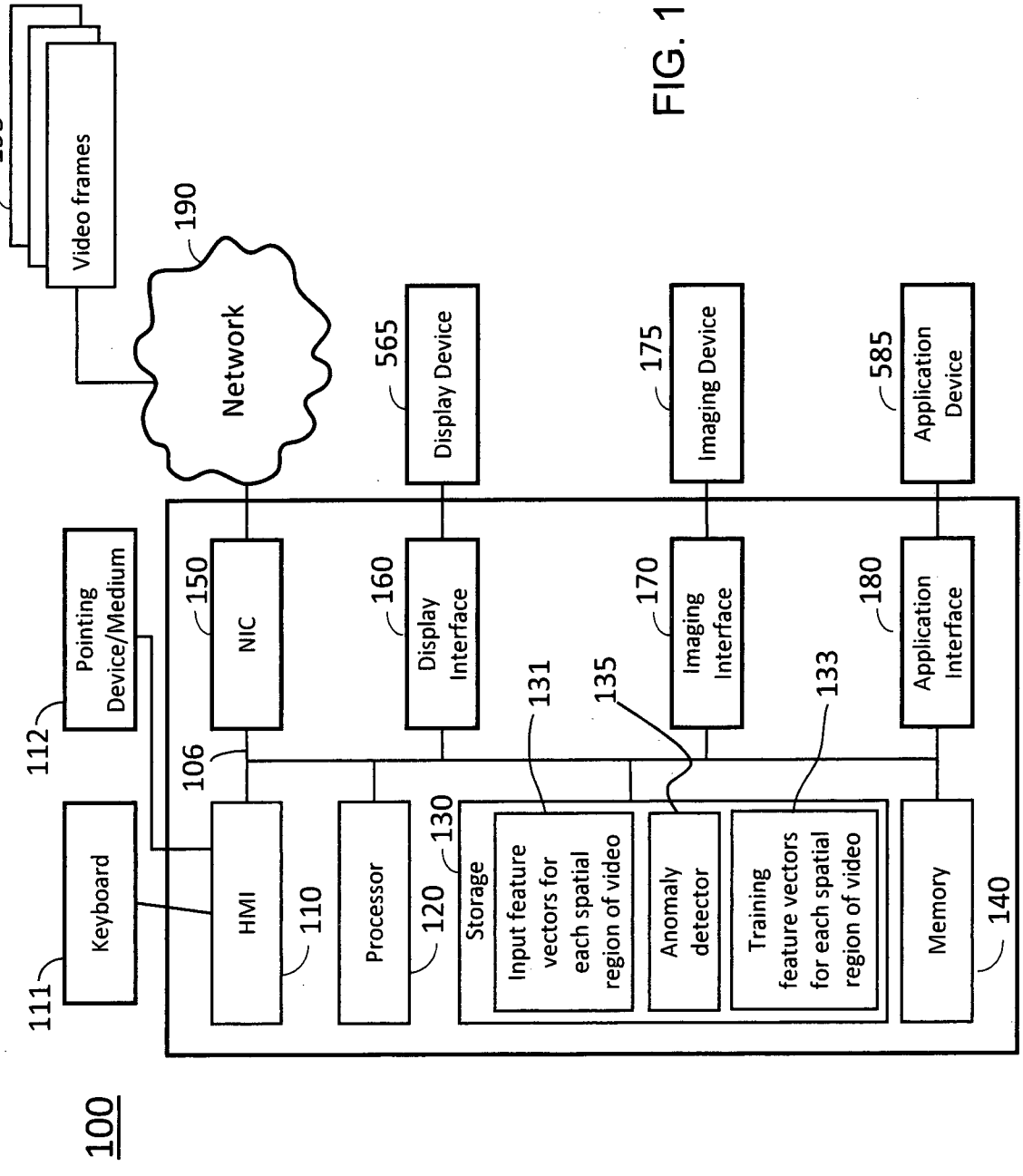
100
111
Keyboard
112
Pointing Device/Medium
HMI
110
106
NIC
150
Network
190
195
Video frames
Processor
120
Storage
130
Input feature vectors for each spatial region of video
131
Anomaly detector
135
Training feature vectors for each spatial region of video
133
Memory
140
Display Interface
160
Display Device
565
Imaging Interface
170
Imaging Device
175
Application Interface
180
Application Device
585

FIG. 1

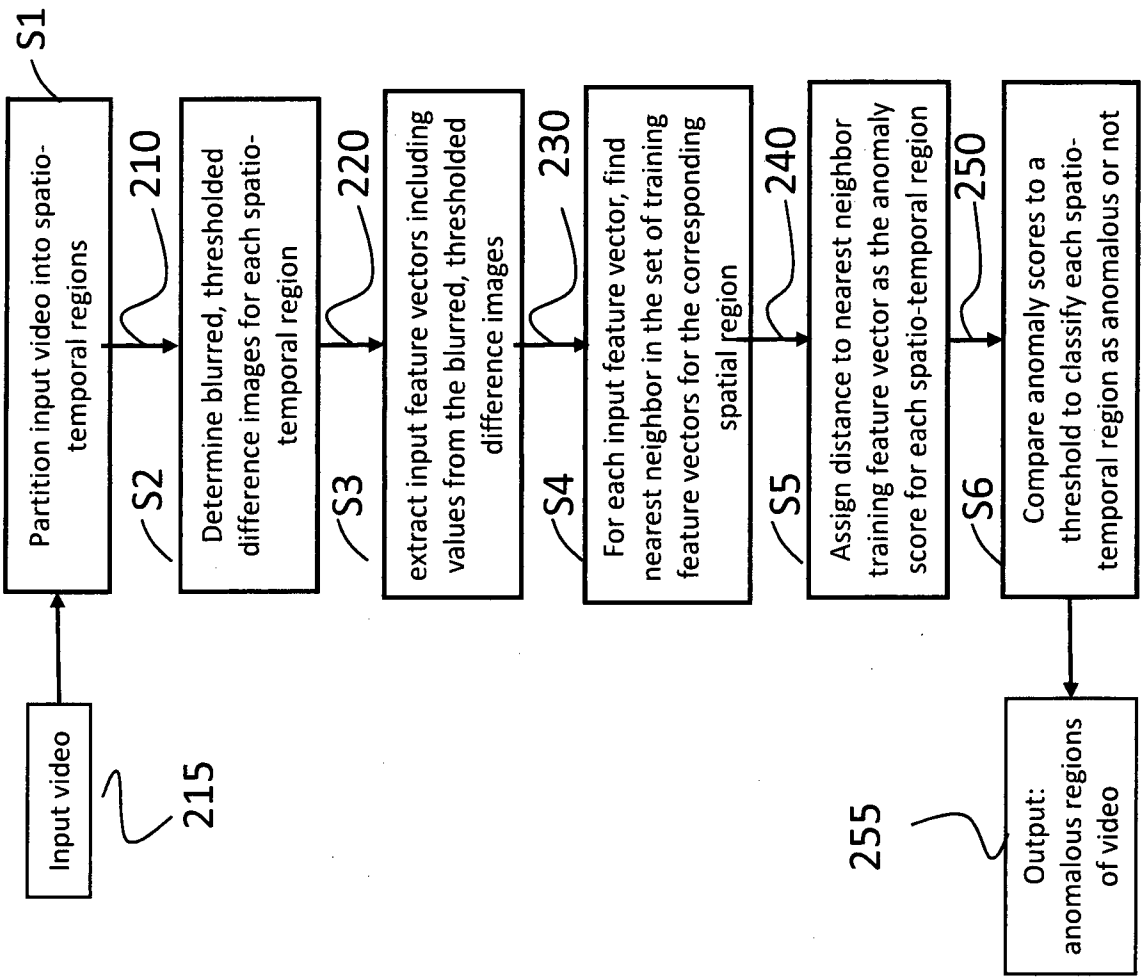
Input video
215
Partition input video into spatio-temporal regions
S1
210
Determine blurred, thresholded difference images for each spatio-temporal region
S2
220
extract input feature vectors including values from the blurred, thresholded difference images
S3
230
For each input feature vector, find nearest neighbor in the set of training feature vectors for the corresponding spatial region
S4
240
Assign distance to nearest neighbor training feature vector as the anomaly score for each spatio-temporal region
S5
250
Compare anomaly scores to a threshold to classify each spatio-temporal region as anomalous or not
S6
255
Output: anomalous regions of video

FIG. 2A

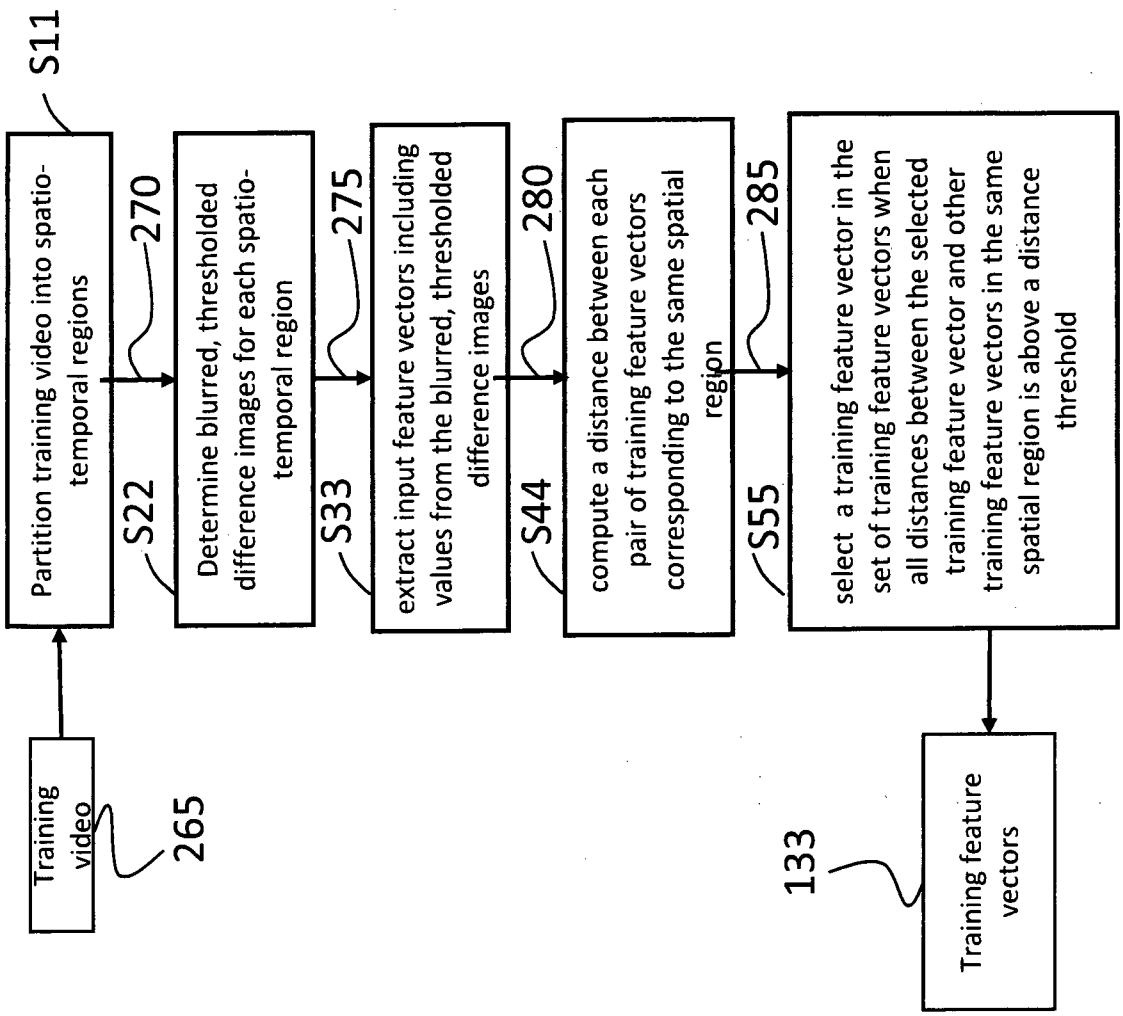
Training video
265
Partition training video into spatio-temporal regions
S11
Determine blurred, thresholded difference images for each spatio-temporal region
S22
270
extract input feature vectors including values from the blurred, thresholded difference images
S33
275
compute a distance between each pair of training feature vectors corresponding to the same spatial region
S44
280
select a training feature vector in the set of training feature vectors when all distances between the selected training feature vector and other training feature vectors in the same spatial region is above a distance threshold
S55
285
Training feature vectors
133

FIG. 2B

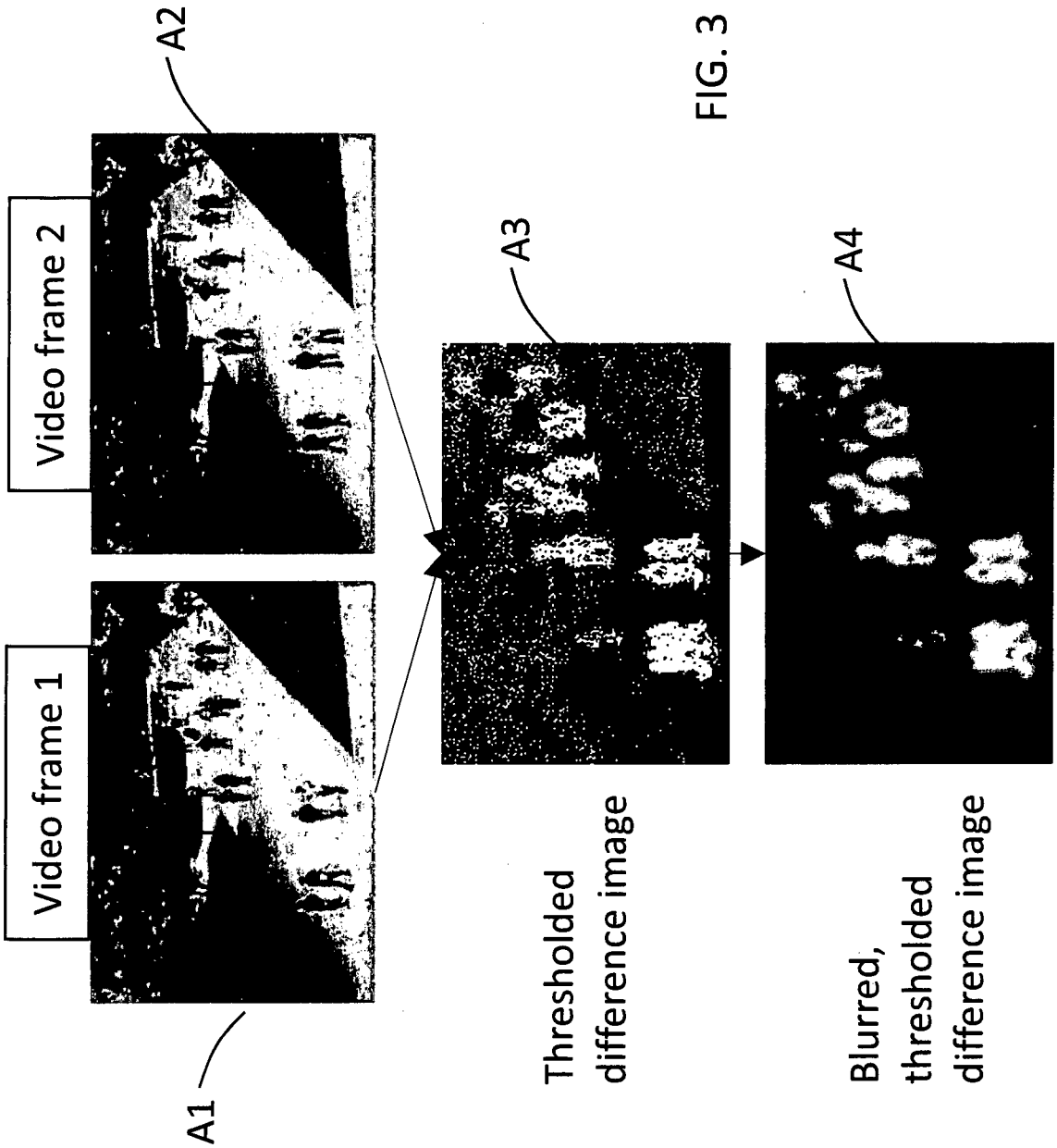
Video frame 1
Video frame 2
A1
A2
A3
A4
Thresholded difference image
Blurred, thresholded difference image

FIG. 3

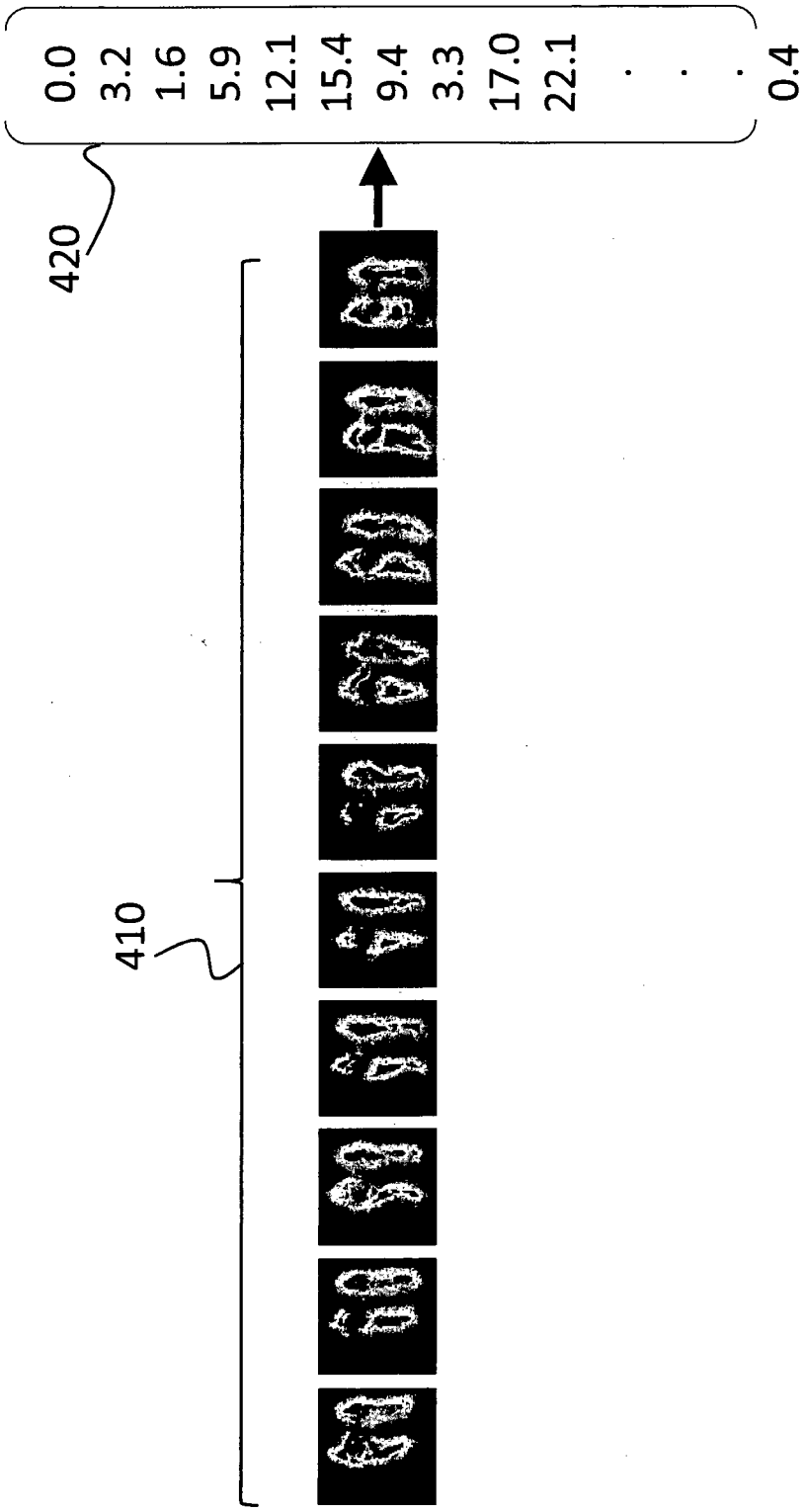

FIG. 4

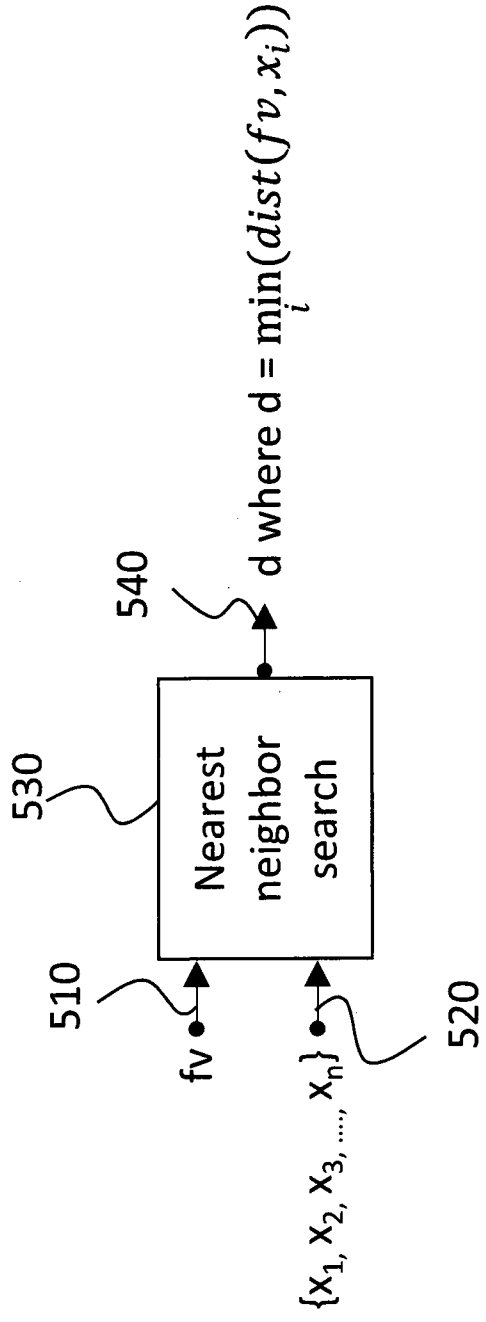
d where d = $\min_i(dist(fv, x_i))$
540
530
Nearest neighbor search
510
fv
$\{x_1, x_2, x_3, \ldots, x_n\}$
520

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2019/002608

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/00
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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TAN HANLIN ET AL: "Fast anomaly detection in traffic surveillance video based on robust sparse optical flow", 2016 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE, 20 March 2016 (2016-03-20), pages 1976-1980, XP032900948, DOI: 10.1109/ICASSP.2016.7472022 [retrieved on 2016-05-18] the whole document ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 April 2019	02/05/2019

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	Koutroumpas, K

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JODOIN P-M ET AL: "Behavior subtraction", VISUAL COMMUNICATIONS AND IMAGE PROCESSING; 20-1-2004 - 20-1-2004; SAN JOSE,, vol. 6822, 1 January 2008 (2008-01-01), pages 68220B-1, XP008098903, DOI: 10.1117/12.770757 ISBN: 978-1-62841-730-2 Section IV -----	1-15