

(51) International Patent Classification:
G06K 9/00 (2006.01) G06N 3/02 (2006.01)(21) International Application Number:
PCT/US2019/043544(22) International Filing Date:
26 July 2019 (26.07.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/722,253 24 August 2018 (24.08.2018) US
16/515,713 18 July 2019 (18.07.2019) US(71) Applicant: NEC LABORATORIES AMERICA, INC.
[US/US]; 4 Independence Way, Suite 200, Princeton, NJ
08540 (US).(72) Inventors: SHARMA, Gaurav; 39620 Potrero Dr.,
Newark, CA 94560 (US). CHANDRAKER, Manmohan;
444 Saratoga Ave., Santa Clara, CA 95050 (US). CHOI,
Jinwoo; 1236 University City Blvd., Blackburg, VA 24060
(US).(74) Agent: BITETTO, James, J.; Tutunjian & Bitetto, P.C.,
401 Broadhollow Road, Suite 302, Melville, NY 11747
(US).(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, 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,
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).

(54) Title: HUMAN ACTION RECOGNITION IN DRONE VIDEOS

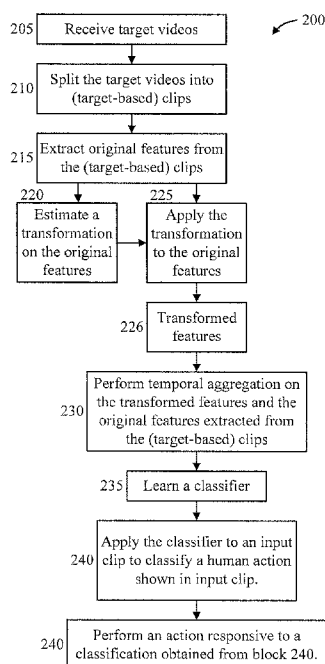


FIG. 2

(57) Abstract: A method is provided for drone- video-based action recognition. The method learns (220) a transformation for each of target video clips taken from a set of target videos, responsive to original features extracted from the target video clips. The transformation corrects differences between a target drone domain corresponding to the target video clips and a source non-drone domain corresponding to source video clips taken from a set of source videos. The method adapts (225) the target to the source domain by applying the transformation to the original features to obtain transformed features for the target video clips. The method converts (230) the original and transformed features of same ones of the target video clips into a single classification feature for each of the target videos. The method classifies (240) a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos.

Published:

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

HUMAN ACTION RECOGNITION IN DRONE VIDEOS

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Provisional Patent Application Serial Number 62/722,253, filed on August 24, 2018, incorporated herein by reference herein its entirety. This application also claims priority to U.S. Non-Provisional Patent Application Serial Number 16/515,713 filed on July 18, 2019, and which is incorporated by reference herein in its entirety.

BACKGROUND

Technical Field

[0002] The present invention relates to information processing and more particularly to human action recognition in drone videos.

Description of the Related Art

[0003] It is desirable to perform human action recognition in videos obtained by drones (hereinafter “drone videos”). Often a sufficient amount of source videos have annotations while target domain videos lack such annotations. Further, such annotations are not readily obtained for the target domain. Hence, there is a need for human action recognition in drone videos capable of overcoming the aforementioned problems.

SUMMARY

[0004] According to an aspect of the present invention, a computer-implemented method is provided for drone-video-based action recognition. The method includes learning, by a hardware processor, a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality

of target video clips. The respective transformation is for correcting differences between a target drone domain corresponding to the plurality of target video clips and a source non-drone domain corresponding to a plurality of source video clips taken from a set of source videos. The method further includes adapting, by the hardware processor, the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the plurality of target video clips. The method also includes converting, by the hardware processor, the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set. The method additionally includes classifying, by the hardware processor, a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

[0005] According to another aspect of the present invention, a computer program product is provided for unsupervised domain adaptation for video classification. The computer program product includes a non-transitory computer readable storage medium having program instructions embodied therewith. The program instructions are executable by a computer to cause the computer to perform a method. The method includes learning, by a hardware processor, a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality of target video clips. The respective transformation is for correcting differences between a target drone domain corresponding to the plurality of target video clips and a source non-drone domain corresponding to a plurality of source video clips taken from a set of source videos. The method further includes adapting, by the hardware processor, the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the

plurality of target video clips. The method also includes converting, by the hardware processor, the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set. The method additionally includes classifying, by the hardware processor, a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

[0006] According to yet another aspect of the present invention, a computer processing system is provided for drone-video-based action recognition. The computer processing system includes a memory for storing program code. The computer processing system further includes a hardware processor for running the program code to learn a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality of target video clips. The respective transformation is for correcting differences between a target drone domain corresponding to the plurality of target video clips and a source non-drone domain corresponding to a plurality of source video clips taken from a set of source videos. The hardware processor further runs the program code to adapt the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the plurality of target video clips. The hardware processor also runs the program code to convert the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set. The processor additionally runs the program code to classify a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

[0007] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0008] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0009] FIG. 1 is a block diagram showing an exemplary processing system, in accordance with an embodiment of the present invention;

[0010] FIG. 2 is a high-level block/flow diagram showing an exemplary training method for unsupervised domain adaptation for video classification, in accordance with an embodiment of the present invention; and

[0011] FIG. 3 is a high-level block/flow diagram showing an exemplary testing (inference) method for unsupervised domain adaptation for video classification, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0012] Embodiments of the present invention are directed to human action recognition in drone videos.

[0013] In an embodiment, methods and systems are provided for human action recognition in drone videos. In an embodiment, the present invention can use a classifier trained on third person videos, and then finetune the classifier on a new annotated dataset of videos taken from drones. In an embodiment, the present invention allows for better learning of the classifier on the drone videos by adding learnable transformations which bring the domain of the drone videos closer to the first domain. A usual scenario is that there is plenty of

annotated data with third person videos, while only a small amount of data is available for drone videos, as there are lesser amounts of videos available and also the annotation cost is high. The proposed method allows learning of classifier which is better adapted to the drone videos and can potentially work with smaller amounts of annotated data.

[0014] In an embodiment, the present invention adds a block in a video classification network, where the block predicts a general transformation to correct or match the drone videos with those of the third person videos. This allows the classifier trained on the third person videos to better adapt to the drone videos. The transformation learning module is added to the traditional human action video classification network and is learnable end-to-end with the annotations available for the videos taken from the drones.

[0015] The present invention allows correction/mitigation of various factors which make it difficult to use a pre-trained classifier from one type of video (e.g., third person videos) with another type of video (e.g., drone videos). The present invention uses a general learnable transformation element which can learn to predict the required transformation for improving recognition in new types of videos.

[0016] In an embodiment, the present invention involves learning a general transformation prediction element as part of a full classifier, where the general transformation prediction element corrects drone videos and makes third person video classifier better adapted to the drone videos.

[0017] FIG. 1 is a block diagram showing an exemplary processing system 100, in accordance with an embodiment of the present invention. The processing system 100 includes a set of processing units (e.g., CPUs) 101, a set of GPUs 102, a set of memory devices 103, a set of communication devices 104, and set of peripherals 105. The CPUs 101 can be single or multi-core CPUs. The GPUs 102 can be single or multi-core GPUs. The one or more memory devices 103 can include caches, RAMs, ROMs, and other memories (flash,

optical, magnetic, etc.). The communication devices 104 can include wireless and/or wired communication devices (e.g., network (e.g., WIFI, etc.) adapters, etc.). The peripherals 105 can include a display device, a user input device, a printer, an imaging device, and so forth. Elements of processing system 100 are connected by one or more buses or networks (collectively denoted by the figure reference numeral 110).

[0018] In an embodiment, memory devices 103 can store specially programmed software modules in order to transform the computer processing system into a special purpose computer configured to implement various aspects of the present invention. In an embodiment, special purpose hardware (e.g., Application Specific Integrated Circuits, and so forth) can be used to implement various aspects of the present invention.

[0019] In an embodiment, the one or more memory devices 103 include an entailment module 103A. In another embodiment, the entailment module 103A can be implemented as special purpose hardware (e.g., an Application Specific Integrated Circuit, and so forth).

[0020] Of course, the processing system 100 may also include other elements (not shown), as readily contemplated by one of skill in the art, as well as omit certain elements. For example, various other input devices and/or output devices can be included in processing system 100, depending upon the particular implementation of the same, as readily understood by one of ordinary skill in the art. For example, various types of wireless and/or wired input and/or output devices can be used. Moreover, additional processors, controllers, memories, and so forth, in various configurations can also be utilized as readily appreciated by one of ordinary skill in the art. These and other variations of the processing system 100 are readily contemplated by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0021] Moreover, it is to be appreciated that various figures as described below with respect to various elements and steps relating to the present invention that may be implemented, in whole or in part, by one or more of the elements of system 100.

[0022] FIG. 2 is a high-level block/flow diagram showing an exemplary training method 200 for unsupervised domain adaptation for video classification, in accordance with an embodiment of the present invention.

[0023] At block 205, receive target videos.

[0024] At block 210, split the target videos into (target-based) clips.

[0025] At block 215, extract original features from the (target-based) clips.

[0026] At block 220, estimate a transformation on the original features.

[0027] At block 225, apply a/the transformation to the original features to obtain transformed features 226. In an embodiment, one or more estimated transformation (per block 220) and/or one or more predetermined transformations can be performed (the latter bypassing the estimating per block 220 in at least one application of the transformation).

[0028] At block 230, perform temporal aggregation on the transformed features and the original features extracted from the (target-based clips). The output of the temporal aggregation is a respective single feature for each of the target videos.

[0029] At block 235, learn a classifier.

[0030] At block 240, apply the classifier to an input clip to classify a human action shown in input clip. The classification is made relative to a set of classes.

[0031] At block 245, perform an action responsive to a classification obtained from block 240. The action can involve, for example, controlling a machine (to shut down, to power up, to turn on lights or open locked doors), to take an avoidance maneuver by a vehicle (e.g., in an Advanced Driver Assistance System (ADAS)), and so forth).

[0032] FIG. 3 is a block diagram showing an exemplary system 300 for unsupervised domain adaptation for video classification, in accordance with an embodiment of the present invention. One or more elements of system 300 can be implemented by one or more elements of system 100.

[0033] In FIG. 3, $v_1, v_2, \dots, v_N$ are the feature vectors of the N clips (the number of clips in source and target video can be different) of the source video. $\Theta_1, \Theta_2, \dots, \Theta_N$ are the parameters of the transformation function T , which acts on the clip features $v_1, v_2, \dots, v_N$ of the target video and gives the transformed features $T_{\Theta_1}(v_1), T_{\Theta_2}(v_2), \dots, T_{\Theta_N}(v_N)$.

[0034] Inputs to the system 300 include input videos 391 (e.g., taken by a drone).

[0035] The system 300 includes a feature extractor 301, a transformation predictor (regressor) 302, a temporal aggregator 303, and a classifier 304.

[0036] The feature extractor 301 is a convolutional neural network (CNN) which takes short clips of the input video and extracts features of the video. This can be any image or video based CNN. For the target video, the extracted features can be output in the form of feature vectors v_1 through v_N . For the source video, the extracted features can be output in the form of feature vectors u_1 through u_m .

[0037] The transformation predictor (regressor) 302 is itself a neural network (or can be a Support Vector Regressor (SVR)) which takes the features extracted from the input clips, and predicts a potentially unique transformation for each of the clips. Since the whole system is learnt, the intuition is that the transformation predictor 302 learns to “correct” the viewpoint and motion differences between the standard third person videos and the current target videos taken from a more challenging platform. In further detail, the transformation predictor 302 predicts the transformation parameters, e.g., if it is a general projective transformation then the parameter matrix is 3×3 with 8 degrees of freedom (DOF). The transformation type is decided by the system designer and can be any of the following: Isometric (3 DOF),

similarity (4DOF), affine (6 DOF) or projective (8 DOF) transformation. Once the transformation is predicted, it is applied to the clip features to output new transformed features which go further in the pipeline. The transformation predictor 302 includes a set of regressors 302A and a set of multipliers 302B used to output the new transformed features. In further detail, the set of regressors 302A take the clip as input and predict the parameters of the transformation to be applied to that particular clip. The parameter matrix is 3x3 with the degrees of freedom depending on the type of parameter decided by the system designer. Once calculated this transformation is applied on the extracted features to obtain the transformed features.

[0038] The temporal aggregator 303 takes the features (the transformed features for the target, and the original features for the source) of the clips of each video and converts them into one single feature for the whole video. This is a many to one operation, i.e., different videos might have different lengths and hence give different number of clips. The temporal aggregator 303 takes all the clips from a video outputs a single feature. In an embodiment, the temporal aggregator 303 can apply a dimension-wise max, or an average operation.

[0039] The classifier 304 then learns to classify the video features into different classes of interest. The classifier 304 or other device (e.g., a hardware processor) can then be used to perform an action responsive to a classification.

[0040] Thus, in an embodiment, the network is first trained on the first type of videos (e.g., third person videos) without the transformation predictor 302. Then the transformation predictor 302 is inserted and the whole network is finetuned to the new types of videos (e.g., drone videos). Once trained, the full network can then be used to make predictions for the drone videos.

[0041] A description will now be given regarding one or more advantages of the present invention, in accordance with one or more embodiments of the present invention.

[0042] Embodiments of the present invention allow for better performance as a result of the transformation based alignment of videos from drones with traditional third person videos. The present invention would also reduce both the cost and deployment time, as lesser amounts of annotations would be required for the drone videos. These and other advantages are readily determined by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0043] The present invention may be a system, a method, and/or a computer program product at any possible technical detail level of integration. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.

[0044] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals *per se*, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating

through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0045] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0046] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as C++ or the like, and conventional procedural programming languages, such as the “C” programming language or similar programming languages. The computer readable program instructions may execute entirely on the user’s computer, partly on the user’s computer, as a stand-alone software package, partly on the user’s computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user’s computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In

some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention.

[0047] Aspects of the present invention are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[0048] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[0049] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other

device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0050] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the blocks may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[0051] Reference in the specification to “one embodiment” or “an embodiment” of the present invention, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well any other variations, appearing in various places throughout the specification are not necessarily all referring to the same embodiment.

[0052] It is to be appreciated that the use of any of the following “/”, “and/or”, and “at least one of”, for example, in the cases of “A/B”, “A and/or B” and “at least one of A and B”, is

intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of both options (A and B). As a further example, in the cases of “A, B, and/or C” and “at least one of A, B, and C”, such phrasing is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of the third listed option (C) only, or the selection of the first and the second listed options (A and B) only, or the selection of the first and third listed options (A and C) only, or the selection of the second and third listed options (B and C) only, or the selection of all three options (A and B and C). This may be extended, as readily apparent by one of ordinary skill in this and related arts, for as many items listed.

[0053] Having described preferred embodiments of a system and method (which are intended to be illustrative and not limiting), it is noted that modifications and variations can be made by persons skilled in the art in light of the above teachings. It is therefore to be understood that changes may be made in the particular embodiments disclosed which are within the scope of the invention as outlined by the appended claims. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A computer-implemented method for drone-video-based action recognition, comprising:

learning (220), by a hardware processor, a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality of target video clips, the respective transformation for correcting differences between a target drone domain corresponding to the plurality of target video clips and a source non-drone domain corresponding to a plurality of source video clips taken from a set of source videos;

adapting (225), by the hardware processor, the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the plurality of target video clips;

converting (230), by the hardware processor, the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set; and

classifying (240), by the hardware processor, a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

2. The computer-implemented method of claim 1, further comprising extracting (215) the original features using a convolutional neural network.

3. The computer-implemented method of claim 1, wherein said learning step (220) learns the respective transformation that corrects for viewpoint differences between the source domain and the target domain.

4. The computer-implemented method of claim 1, wherein said learning step (220) learns the respective transformation that corrects for motion differences between the source domain and the target domain.

5. The computer-implemented method of claim 1, wherein said converting step (230) is performed on the original features and the transformed features using a dimension-wise maximum function.

6. The computer-implemented method of claim 1, wherein said converting step (230) is performed the original features and the transformed features using an averaging function.

7. The computer-implemented method of claim 1, wherein the hardware processor uses a Supper Vector Regressor to predict parameters of the respective transformation.

8. The computer-implemented method of claim 1, further comprising capturing, by a drone, the target videos in the set.

9. The computer-implemented method of claim 1, wherein said adapting step finetunes a neural network used said classifying step to align the target drone domain to the source non-drone domain.

10. A computer program product for unsupervised domain adaptation for video classification, the computer program product comprising a non-transitory computer readable storage medium having program instructions embodied therewith, the program instructions executable by a computer to cause the computer to perform a method comprising:

learning (220), by a hardware processor, a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality of target video clips, the respective transformation for correcting differences between a target domain corresponding to the plurality of target video clips and a source domain corresponding to a plurality of source video clips taken from a set of source videos;

adapting (225), by the hardware processor, the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the plurality of target video clips;

converting (230), by the hardware processor, the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set; and

classifying (240), by the hardware processor, a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

11. The computer-implemented method of claim 10, wherein the method further comprises extracting (215) the original features using a convolutional neural network.

12. The computer-implemented method of claim 10, wherein said learning step (220) learns the respective transformation that corrects for viewpoint differences between the source domain and the target domain.

13. The computer-implemented method of claim 10, wherein said learning step (220) learns the respective transformation that corrects for motion differences between the source domain and the target domain.

14. The computer-implemented method of claim 10, wherein said converting step (230) is performed on the original features and the transformed features using a dimension-wise maximum function.

15. The computer-implemented method of claim 10, wherein said converting step (230) is performed the original features and the transformed features using an averaging function.

16. The computer-implemented method of claim 10, wherein the hardware processor uses a Supper Vector Regressor to predict parameters of the respective transformation.

17. The computer-implemented method of claim 10, wherein the method further comprises capturing, by a drone, the target videos in the set.

18. The computer-implemented method of claim 10, wherein said adapting step finetunes a neural network used said classifying step to align the target drone domain to the source non-drone domain.

19. A computer processing system for drone-video-based action recognition, comprising:

a memory (103) for storing program code; and

a hardware processor (101) for running the program code to

learn a respective transformation for each of a plurality of target video clips taken from a set of target videos, responsive to original features extracted from the plurality of target video clips, the respective transformation for correcting differences between a target drone domain corresponding to the plurality of target video clips and a source non-drone domain corresponding to a plurality of source video clips taken from a set of source videos;

adapt the target domain to the source domain by applying the respective transformation to the original features extracted from the plurality of target video clips to obtain transformed features for the plurality of target video clips;

convert the original features and the transformed features of same ones of the plurality of target video clips into a single classification feature for each of the target videos in the set; and

classify a human action in a new target video relative to the set of source videos using the single classification feature for each of the target videos in the set.

20. The computer processing system of claim 19, wherein the hardware processor (101) uses a Support Vector Regressor to predict parameters of the respective transformation.

100→

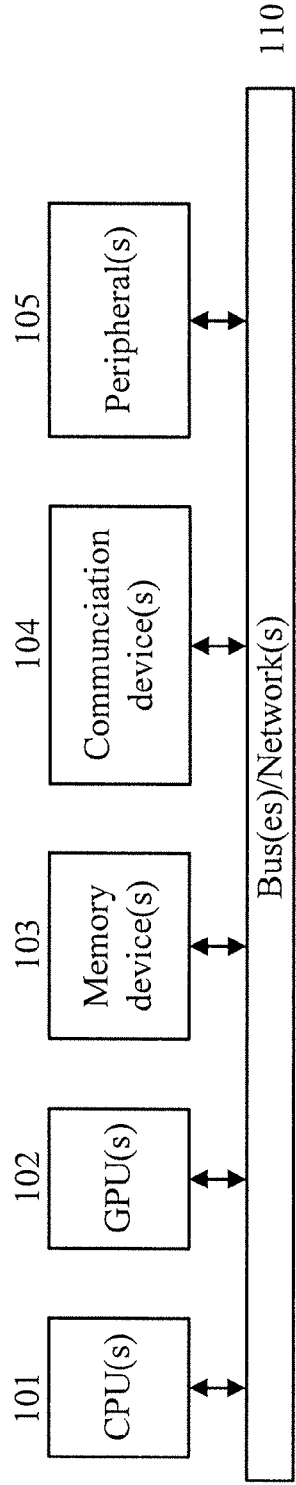


FIG. 1

2/3

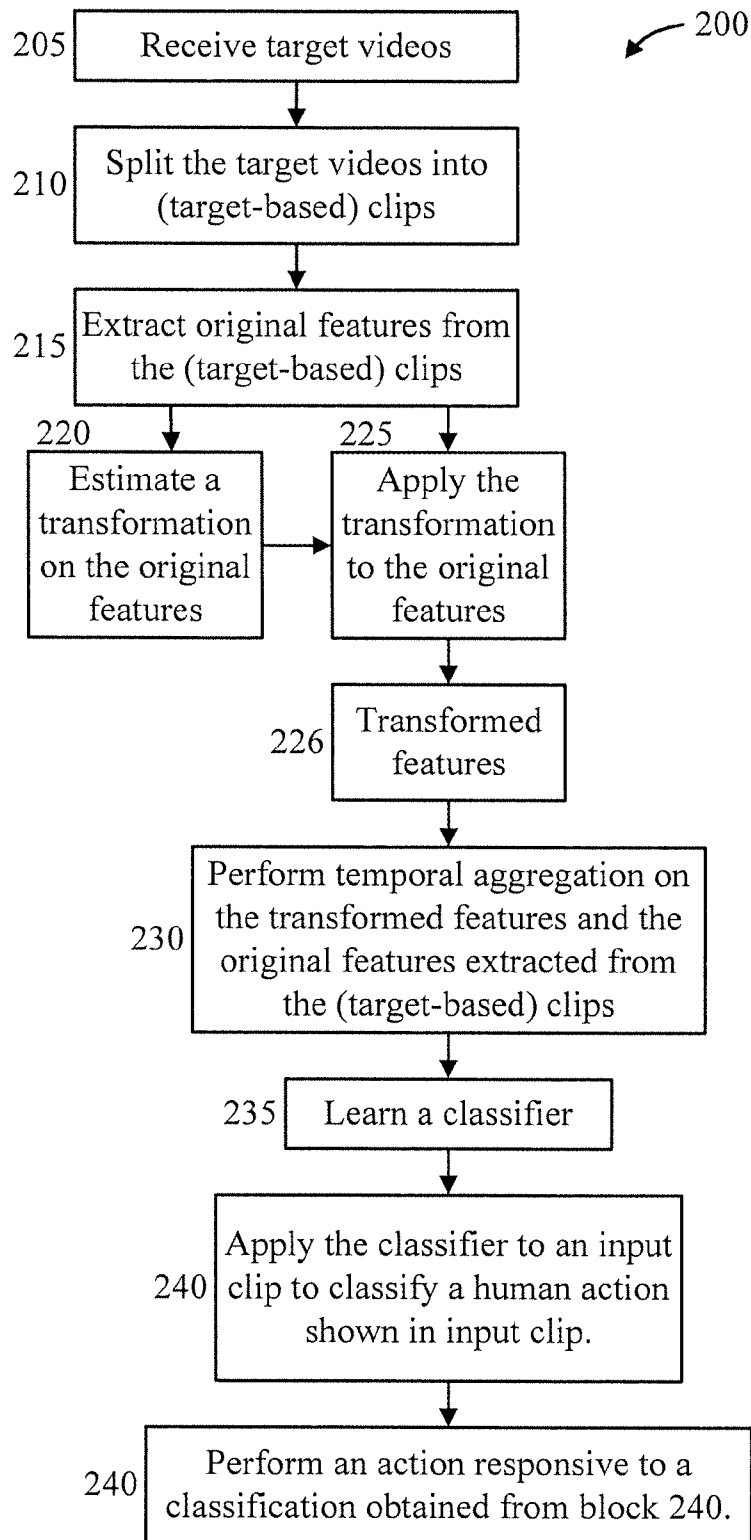


FIG. 2

3/3

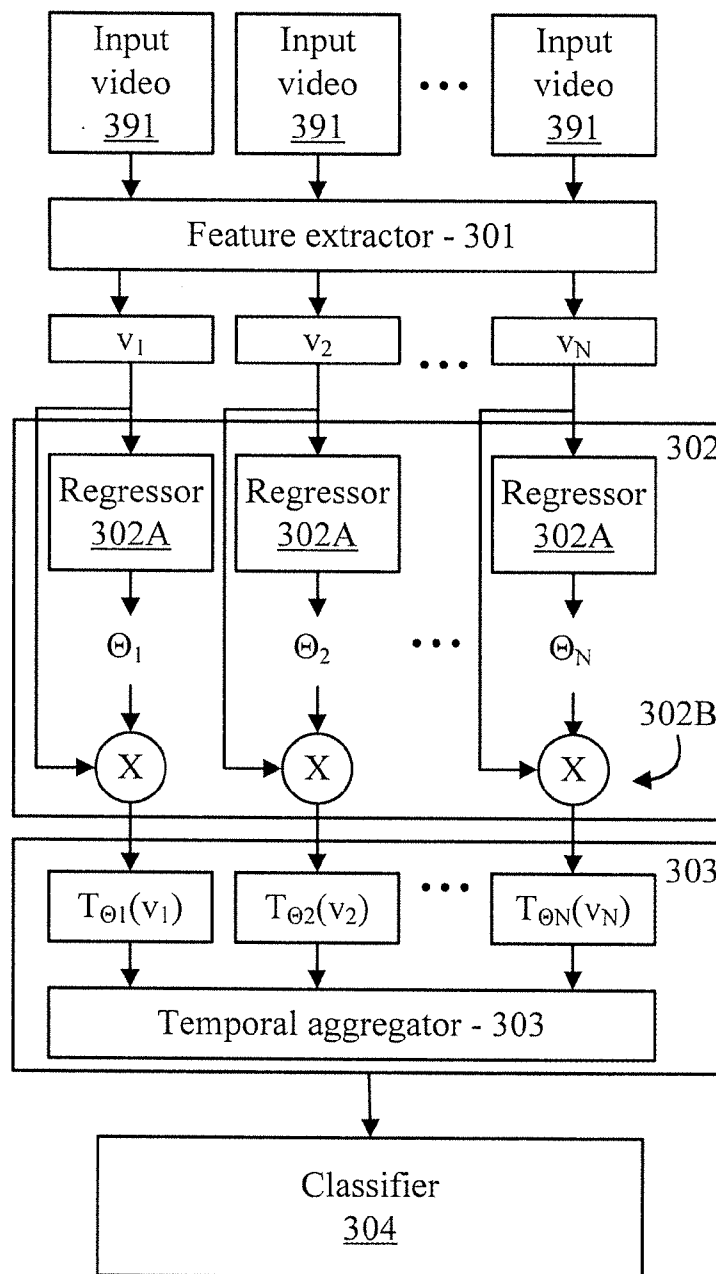


FIG. 3

A. CLASSIFICATION OF SUBJECT MATTER**G06K 9/00(2006.01)i, G06N 3/02(2006.01)i**

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 9/00; G06K 9/46; G06Q 50/22; G06N 3/02

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) & Keywords: drone, action, recognition, transformation, single classification feature, original feature, transformation feature

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	AMARJOT SINGH et al., `Eye in the Sky: Real-time Drone Surveillance System (DSS) for Violent Individuals Identification using ScatterNet Hybrid Deep Learning Network`, IEEE Computer Vision and Pattern Recognition (CVPR), pages 1742-1750, 2018.06.03 < https://arxiv.org/abs/1806.00746 > See pages 1742-1750 and figures 1-3.	1-20
A	HOSSEIN RAHMANI et al., `Learning a Deep Model for Human action Recognition from Novel Viewpoints`, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 40, Issue: 3, pages 667-681, 2018.03.01 < https://ieeexplore.ieee.org/document/7893732 > See pages 667-681 and figure 1.	1-20
A	MATIJA RADOVIC et al., `Object Recognition in Aerial Images Using Convolutional Neural Networks`, J. Imaging 2017, Volume: 3, Issue: 2, pages 1-9, 2017.06.14 < https://www.mdpi.com/2313-433X/3/2/21 > See pages 1-9.	1-20
A	KR 10-2018-0049786 A (SAMSUNG ELECTRONICS CO., LTD.) 11 May 2018 See paragraphs [0037]-[0080] and figures 1a, 1b.	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

"D" document cited by the applicant in the international application

"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

15 November 2019 (15.11.2019)

Date of mailing of the international search report

18 November 2019 (18.11.2019)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

KANG, Sung Chul

Telephone No. +82-42-481-8405



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2019/043544

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017-0109582 A1 (DISNEY ENTERPRISES, INC.) 20 April 2017 See claims 1-7 and figures 1-3.	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2019/043544

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
KR 10-2018-0049786 A	11/05/2018	CN 109906457 A EP 3532990 A1 US 2018-0121732 A1 WO 2018-084577 A1	18/06/2019 04/09/2019 03/05/2018 11/05/2018
US 2017-0109582 A1	20/04/2017	US 9805264 B2	31/10/2017