



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

(21) International Application Number:
PCT/CN2016/078982

(22) International Filing Date:
11 April 2016 (11.04.2016)

(25) Filing Language: English

(26) Publication Language: English

(72) Inventor; and

(71) Applicant : **WANG, Xiaogang** [CN/CN]; Room 415, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (CN).

(72) Inventors: **SHAO, Jing**; Room 304, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (CN). **LOY, Chen-Change**; Room 717, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong Shatin, N.T., Hong Kong (CN). **KANG, Kai**; Room 304, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong Shatin, N.T., Hong Kong (CN).

(74) Agent: **INSIGHT INTELLECTUAL PROPERTY LIMITED**; 19A, Tower A, InDo Building, No. 48A Zhichun Road, Haidian District, Beijing 100098 (CN).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

(54) Title: METHOD AND SYSTEM FOR OBJECT TRACKING

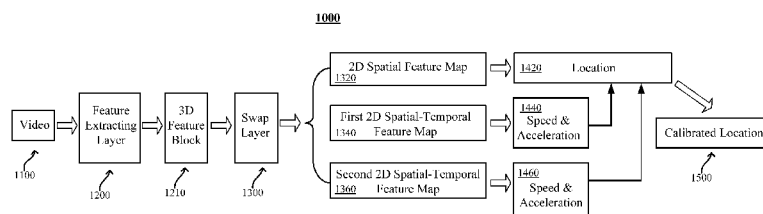


Fig. 1

(57) Abstract: The disclosures relate to a method and system for tracking a target object in a video. The method comprises: extracting, from the video, a 3-dimension (3D) feature block containing the target object; decomposing the extracted 3D feature block into a 2-dimension (2D) spatial feature map containing spatial information of the target object and a 2D spatial-temporal feature map containing spatial-temporal information of the target object; estimating, in the 2D spatial feature map, a location of the target object; determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object; calibrating the estimated location of the target object according to the determined speed and acceleration; and tracking the target object in the video according to the calibrated location.

METHOD AND SYSTEM FOR OBJECT TRACKING

Technical Field

[0001] The disclosures relate to a method and a system for tracking a target object in a video.

Background

[0002] It is a crucial task to track a target object in a video. Generally, the target object (e.g., a group or an individual) should first be recognized from each frame of the video. Then a tracking is performed to capture movements of objects separately. Therefore, a location of the target object and dynamic features (e.g., speed and acceleration) thereof are detected in separate tasks, which limits accuracy of object tracking.

[0003] A lot of learning-based methods are developed for the object tracking. Convolutional Neural Network (CNN), as a successful deep learning model applied in the object tracking, has demonstrated superior performances in speed and accuracy. It is desired to develop a CNN to enhance the accuracy of the object tracking.

Summary

[0004] The following presents a simplified summary of the disclosure in order to provide a basic understanding of some aspects of the disclosure. This summary is not an extensive overview of the disclosure. This summary neither identifies key or critical elements of the disclosure nor delineates any scope of particular embodiments of the disclosure, or any scope of the claims. Its sole purpose is to present some concepts of the disclosure in a simplified form as a prelude to the more detailed description that is presented later.

[0005] In order to address, at least partially, one of the above issues, a method and a system for tracking a target object in a video is provided. The method comprises: extracting, from the

video, a 3-dimension (3D) feature block containing the target object; decomposing the extracted 3D feature block into a 2-dimension (2D) spatial feature map containing spatial information of the target object and a 2D spatial-temporal feature map containing spatial-temporal information of the target object; estimating, in the 2D spatial feature map, a location of the target object; determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object; calibrating the estimated location of the target object according to the determined speed and acceleration; and tracking the target object in the video according to the calibrated location.

[0006] In one embodiment of the present application, the 2D spatial feature map may extend in a first spatial direction and a second spatial direction intersecting with the first spatial direction.

[0007] In one embodiment of the present application, the 2D spatial-temporal feature map may comprise: a first 2D spatial-temporal feature map extending in the first spatial direction and a temporal direction and comprising components of the speed and the acceleration of the target object in the first spatial direction; and a second 2D spatial-temporal feature map extending in the second spatial direction and the temporal direction and comprising components of the speed and the acceleration of the target object in the second spatial direction.

[0008] In one embodiment of the present application, the method may comprise providing a CNN comprising a feature extracting layer, wherein the extracting may comprise: filtering, in the feature extracting layer, each frame of the video to obtain first feature maps; evaluating overlap degrees and similarity degrees between the first feature maps and a preset image containing a feature of interest (FOI) of the target object; and selecting, from the first feature maps, a second feature map according to the overlap degrees and the similarity degrees, wherein the second feature map only contains the FOI of the target object; and combining the selected second feature map over each frame of the video together to construct the 3D feature block. In an alternative implementation, the evaluating comprises comparing the first feature

maps with a binary mask generated from the preset image.

[0009] In one embodiment of the present application, the CNN may further comprise a swap layer coupled to the feature extracting layer, and wherein the decomposing may comprise: receiving the 3D feature block from the feature extracting layer; disabling data of the received feature block in the temporal direction to obtain the 2D spatial feature map; and disabling data of the received feature block in one of the first spatial direction and the second spatial direction to obtain the 2D spatial-temporal feature map.

[0010] In one embodiment of the present application, the CNN may further comprise a 2D spatial feature extracting layer coupled to the swap layer, and wherein the estimating may comprise: receiving the 2D spatial feature map from the swap layer; enhancing discrepancies of the FOI in the 2D spatial feature map; recognizing the target object according to the enhanced FOI; and estimating the location of the recognized target object.

[0011] In one embodiment of the present application, the CNN may further comprise a 2D spatial-temporal feature extracting layer coupled to the swap layer and parallel to the 2D spatial feature extracting layer, and wherein the determining may comprise: receiving the 2D spatial-temporal feature map from the swap layer; enhancing discrepancies of the FOI in the 2D spatial-temporal feature map; recognizing the target object according to the enhanced FOI; and performing a derivation operation on the recognized target object in the temporal direction, to determine the speed and the acceleration of the target object.

[0012] In one embodiment of the present application, the method may comprise: training the feature extracting layer independently; and training, based on the trained feature extracting layer, the 2D spatial feature extracting layer and the 2D spatial-temporal feature extracting layer separately.

[0013] In another aspect, a system for tracking a target object in a video is provided. The system may comprise: a memory that stores executable components; and a processor

electrically coupled to the memory to execute the executable components. The executable components is configured for: extracting, from the video, a 3D feature block containing the target object; decomposing the extracted 3D feature block into: a 2D spatial feature map containing spatial information of the target object; and a 2D spatial-temporal feature map containing spatial-temporal information of the target object; estimating, in the 2D spatial feature map, a location of the target object; determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object; calibrating the estimated location of the target object according to the determined speed and acceleration; and tracking the target object in the video according to the calibrated location.

[0014] In yet another aspect, a system for tracking a target object in a video is provided. The system may comprise: a feature extractor configured for extracting, from the video, a 3D feature block containing the target object; a decomposer configured for decomposing the extracted 3D feature block into a 2D spatial feature map containing spatial information of the target object and a 2D spatial-temporal feature map containing spatial-temporal information of the target object; a locator configured for estimating, in the 2D spatial feature map, a location of the target object; a motion detector configured for determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object; a calibrator configured for calibrating the estimated location of the target object according to the determined speed and acceleration; and a tracker configured for tracking the target object in the video according to the calibrated location.

Brief Description of the Drawing

[0015] Exemplary non-limiting embodiments of the present application are described below with reference to the attached drawings. The drawings are illustrative and generally not to an exact scale. The same or similar elements on different figures are referenced with the same reference numbers.

[0016] Fig. 1 illustrates an exemplary system for tracking a target object in a video

according to one embodiment of the present application;

[0017] Fig. 2 is a flow chart illustrating a process for tracking a target object in a video according to one embodiment of the present application; and

[0018] Fig. 3 illustrates an exemplary system for providing attribute information of a video according to one embodiment of the present application.

Detailed Description

[0019] Reference will now be made in detail to some specific embodiments of the present application contemplated by the inventors for carrying out the present application. Examples of these specific embodiments are illustrated in the accompanying drawings. While the present application is described in conjunction with these specific embodiments, it will be appreciated by one skilled in the art that it is not intended to limit the present application to the described embodiments. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present application. The present application may be practiced without some or all of these specific details. In other instances, well-known process operations have not been described in detail in order not to unnecessarily obscure the present application.

[0020] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the present application. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising" when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0021] An exemplary system 1000 for tracking the object is illustrated in Fig. 1. Hereinafter,

the systems and the processes for object tracking will be described under frameworks of CNNs for the purpose of illustration. A video 1100 containing a target object is input to a feature extracting layer 1200. The feature extracting layer 1200 may be implemented in various CNN models, such as AlexNet, Clarifai, Overfeat, GoogleNet, and Visual Geometry Group (VGG). As would be appreciated by one skilled in the art, the feature extracting layer 1200 may be embodied in a plurality of layers. Accordingly, features in lower layers capture low-level information, while those in higher layers can represent more abstract concepts. A plurality of filtering operations may be applied to the video 1100 to extract 3D feature blocks 1210. The 3D feature blocks 1210 may extend in, for example but not limited to, a horizontal direction (denoted as X), a vertical direction (denoted as Y), and a temporal direction (denoted as T).

[0022] It is admitted that it will cost a lot to process the 3D feature blocks in a single convoluting operation. As a consequence, the 3D feature blocks 1210 may be decomposed into one or more (for example, three) 2D slices in a swap layer 1300. XY-slices (i.e., 2D spatial feature map 1320) of the 3D feature blocks 1210 may represent feature maps as traditionally extracted from frames of the video 1100 via the filtering operations stated above. Therefore, XY-slices of the 3D feature blocks 1210 contain spatial information of the target object. Accordingly, XT-slices (i.e., first 2D spatial-temporal feature maps 1340) and YT-slices (i.e., second 2D spatial-temporal feature maps 1360) contain spatial-temporal information of the target object since they both extend in a spatial direction and a temporal direction. The decomposing operation may be implemented by enabling data in two of the three directions and disabling data in the remaining direction.

[0023] As has been discussed above, the 2D spatial feature maps 1320 can be regarded as feature maps extracted from a frame of the video 1100. The 2D spatial feature maps 1320 may comprise various types of semantic meaningful representations of various objects and backgrounds. In an alternative implementation, the 2D spatial feature maps 1320 may be further filtered to enhance representation capacities (discrepancies) thereof. With the

enhanced discrepancies, the target object contained in each frame of the video can be recognized from other objects and backgrounds. Accordingly, a first set of locations 1420 can be estimated based on the recognized object.

[0024] The 2D spatial-temporal feature maps 1340 and 1360 represent profiles of the video along a spatial direction. The 2D spatial-temporal feature maps 1340 and 1360 may also comprise various types of semantic meaningful representations of various objects and backgrounds. In an alternative implementation, the 2D spatial-temporal feature maps 1340 and 1360 may also be further filtered to enhance representation capacities (discrepancies) thereof. With the enhanced discrepancies, the target object contained in each frame of the video can be recognized from other objects and backgrounds. Although 2D spatial-temporal feature maps 1340 and 1360 have less spatial representations as those of the 2D spatial feature maps 1320, the 2D spatial-temporal feature maps 1340 and 1360 have additional information related to time. Therefore, dynamic information (e.g., speed and acceleration 1440 and 1460), can be derived from the 2D spatial-temporal feature maps 1340 and 1360 by, for example but not limited to, a derivation operation.

[0025] It is crucial to obtain the dynamic information of recognized targets in the field of object tracking. For example, the dynamic information may be utilized to predict the motion of the target object in subsequent frames of the video. In applications, the speed and the acceleration in the X direction (denoted as V_X and A_X , respectively) may be derived from the XT-slices (i.e., first 2D spatial-temporal feature maps 1340), and the speed and the acceleration in the Y direction (denoted as V_Y and A_Y , respectively) may be derived from the YT-slices (i.e., first 2D spatial-temporal feature maps 1340). At least one of the dynamic information of V_X , A_X , V_Y , and A_Y may be utilized to calibrate the first set of locations 1420 estimated previously in the 2D spatial feature maps. With the calibrated locations 1500, it is possible to track the object in a higher accuracy.

[0026] Fig. 2 illustrates the above-mentioned process 2000 for tracking a target object in a video. A 3D feature block is extracted from an input video at step S2010. The extracted 3D

feature block may be decomposed into a 2D spatial feature map and a 2D spatial-temporal feature map at step S2020. In an alternative implementation, the 2D spatial feature map extends in a first spatial direction and a second spatial direction intersecting with the first spatial direction. In such a case, the 2D spatial-temporal feature map may comprise a first 2D spatial-temporal feature map extending in the first spatial direction and a temporal direction and a second 2D spatial-temporal feature map extending in the second spatial direction and the temporal direction. At step S2030, a location of a recognized object may be estimated in the 2D spatial feature map. At step S2040, a speed and an acceleration may be determined in the 2D spatial-temporal feature map. With the determined speed and acceleration, the location estimate previously may be calibrated at step S2050. At step S2060, the calibrated location may be used to perform an object tracking task in the video.

[0027] As has been described above, a plurality of filtering operations are applied to the video to extract 3D feature blocks, wherein each of the 3D feature blocks represents a particular feature in the video, such as a human head, a leaf, a crowd scene, etc. In practical applications, not all the 3D feature blocks are crucial to recognize the target object. For example, in the case that the target object is a young lady, 3D feature blocks representing irrelevant features (such as roof corners, water waves) may be ignored in subsequent operations to relieve computing burden. On the contrary, the 3D feature blocks representing relevant features (such as hairs, shoes), which is referred to as features of interest (FOIs), may be reserved. This process is called a pruning operation. The key point in the pruning operation resides in how to assess the relevance between the candidate feature and the FOI. In an alternative embodiment, first, the frame is forward to the feature extracting layer to convolute with a first set of filters, which results in a first set of feature maps. Then, the relevance is assessed by investigating the spatial distribution of the first set of feature maps over binary masks containing FOIs of the target object, wherein the binary masks are prepared according to a preset validation set of images as would be appreciated by one skilled in the art.

[0028] In the assessment, two benchmarks are adopted, which are referred to as an affinity

score and a conspicuous score. The affinity scores measure the overlap degrees between the first set of feature maps and the binary masks. For the i -th binary mask $\mathbf{S}_i$ and the n -th feature map in the first set of the feature maps $\mathbf{F}_i^n$, the affinity score α_i^n is expressed by rule of:

$$\alpha_i^n = \|\mathbf{1}_{[\mathbf{F}_i^n > 0]} \bullet \mathbf{S}_i\|_1 / \|\mathbf{S}_i\|_1 \quad (1)$$

where $\mathbf{1}_{[\cdot]}$ is an indicator function that returns 1 when its input is true, and $\bullet$ denotes an element-wise multiplication operation. The conspicuous scores measure the similarity degrees between the first set of feature maps and the binary masks. For the i -th binary mask $\mathbf{S}_i$ and the n -th feature map in the first set of the feature maps $\mathbf{F}_i^n$, the conspicuous score κ_i^n is expressed by rule of:

$$\kappa_i^n = \|\mathbf{F}_i^n \bullet \mathbf{S}_i\|_1 / \|\mathbf{F}_i^n\|_1 \quad (2)$$

[0029] Then a histogram $\mathbf{H}$ is constructed, with respect to the n -th feature map, to assess the relevance between n -th feature map and the FOIs contained in the binary masks. The histogram $\mathbf{H}$ is summed over each of the binary mask subscripts $i \in [1, m]$ according to the following logic:

$$Hn(i+1) = Hn(i) + \mathbf{1}_{[\alpha_i^n > \tau_\alpha \cup \kappa_i^n > \tau_\kappa]}, \quad i++, \quad i \in [1, m] \quad (3)$$

The equation (3) expresses that if the affinity score α_i^n is larger than a predetermined threshold τ_α or the conspicuous score κ_i^n is larger than a predetermined threshold τ_κ , the histogram $\mathbf{H}$ is summed by 1. The histogram $\mathbf{Hn}$ will return a certain value after a sum operation over each of the binary mask. The histogram $\mathbf{Hn}$ for each feature map in the first set of feature maps (i.e., for each filter in the first set of filters) is calculated in a same manner and be sorted in a descending order. The number of FOIs may be manually set. If the number of FOIs is set to 10, the feature maps with largest 10 $\mathbf{Hn}$ are selected from the first set of

feature maps to constitute a second set of feature maps. In such a case, 10 3D feature blocks are constructed by combining the selected second set of feature maps over each frame of the video together.

[0030] Fig. 3 illustrates an exemplary system 3000 for providing attribute information of a video according to another embodiment of the present application. A video 3100 containing a target object may be input into a feature extracting layer 3200. A 3D feature block 3210 comprising a FOI of the target object may be extracted from the feature extracting layer 3200 and experience a decomposing operation in a swap layer 3300 to be decomposed into a 2D spatial feature map 3320, a first 2D spatial-temporal feature map 3340, and a second spatial-temporal feature map 3360. The 2D spatial feature map 3320 may contain spatial information of the target object, such as a location of the target object in a frame of the video. The first 2D spatial-temporal feature map 3340 and the second spatial-temporal feature map 3360 may contain spatial-temporal information of the target object, such as speed and acceleration of the target object. The 2D spatial feature map 3320, the first 2D spatial-temporal feature map 3340, and the second spatial-temporal feature map 3360 may be forwarded to a 2D spatial feature branch 3420, a first 2D spatial-temporal feature branch 3440, and a second spatial-temporal feature branch 3460, respectively to further enhance semantic representations (discrepancies) thereof. The 2D feature maps with enhanced semantic representations may be combined in a concatenation layer 3500 to construct a 3D feature maps with spatial and temporal information of the target object. A classifier 3600, such as a SVM, may be applied to the constructed 3D feature maps to provide the attribute information (location, speed, acceleration, etc) of the target object.

[0031] In an alternative implement, the feature extracting layer 3200 may be pre-trained to give out well-learned a 3D feature block. Then the 2D spatial feature branch 3420, the first 2D spatial-temporal feature branch 3440, and the second spatial-temporal feature branch 3460 may be trained separately based on the well-learned 3D feature block. Then the classifier 3600, such as a SVM, may be trained individually based on the well-trained feature extracting

layer 3200, the 2D spatial feature branch 3420, the first 2D spatial-temporal feature branch 3440, and the second spatial-temporal feature branch 3460. This training process enables a high learning efficiency of the whole system 3000.

[0032] As will be appreciated by one skilled in the art, the present application may be embodied as a system, a method or a computer program product. Accordingly, the present application may take the form of an entirely hardware embodiment and hardware aspects that may all generally be referred to herein as a "unit", "circuit", "module", or "system". Much of the functionality and many of the principles when implemented, are best supported with or integrated circuits (ICs), for example but not limited to, a digital signal processor and software therefore or application specific ICs. It is expected that one of ordinary skill, notwithstanding possibly significant effort and many design choices motivated by, for example, available time, current technology, and economic considerations, when guided by the concepts and principles disclosed herein will be readily capable of generating ICs with minimal experimentation. Therefore, in the interest of brevity and minimization of any risk of obscuring the principles and concepts according to the present application, further discussion of such software and ICs, if any, will be limited to the essentials with respect to the principles and concepts used by the preferred embodiments. In addition, the present application may take the form of an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software. For example, the system may comprise a memory that stores executable components and a processor, electrically coupled to the memory to execute the executable components to perform operations of the system, as discussed in reference to Figs. 1-3. Furthermore, the present application may take the form of a computer program product embodied in any tangible medium of expression having computer-usable program code embodied in the medium.

What is claimed is:

1. A method for tracking a target object in a video, comprising:

extracting, from the video, a 3-dimension (3D) feature block containing the target object;
decomposing the extracted 3D feature block into:

a 2-dimension (2D) spatial feature map containing spatial information of the target object; and

a 2D spatial-temporal feature map containing spatial-temporal information of the target object;

estimating, in the 2D spatial feature map, a location of the target object;

determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object;

calibrating the estimated location of the target object according to the determined speed and acceleration; and

tracking the target object in the video according to the calibrated location.

2. The method of claim 1,

wherein the 2D spatial feature map extends in:

a first spatial direction; and

a second spatial direction intersecting with the first spatial direction; and

wherein the 2D spatial-temporal feature map comprises:

a first 2D spatial-temporal feature map extending in the first spatial direction and a temporal direction and comprising components of the speed and the acceleration of the target object in the first spatial direction; and

a second 2D spatial-temporal feature map extending in the second spatial direction and the temporal direction and comprising components of the speed and the acceleration of the target object in the second spatial direction.

3. The method of claim 2, further comprising providing a Convolutional Neural Network (CNN) comprising a feature extracting layer, wherein the extracting comprises:

filtering, in the feature extracting layer, each frame of the video to obtain first feature maps;

evaluating overlap degrees and similarity degrees between the first feature maps and a preset image containing a feature of interest (FOI) of the target object; and

selecting, from the first feature maps, a second feature map according to the overlap degrees and the similarity degrees, wherein the second feature map only contains the FOI of the target object; and

combining the selected second feature map over each frame of the video together to construct the 3D feature block.

4. The method of claim 3, wherein the CNN further comprises a swap layer coupled to the feature extracting layer, and wherein the decomposing comprises:

receiving the 3D feature block from the feature extracting layer;

disabling data of the received feature block in the temporal direction to obtain the 2D spatial feature map; and

disabling data of the received feature block in one of the first spatial direction and the second spatial direction to obtain the 2D spatial-temporal feature map.

5. The method of claim 4, wherein the CNN further comprises a 2D spatial feature extracting layer coupled to the swap layer, and wherein the estimating comprises:

receiving the 2D spatial feature map from the swap layer;

enhancing discrepancies of the FOI in the 2D spatial feature map;

recognizing the target object according to the enhanced FOI; and

estimating the location of the recognized target object.

6. The method of claim 5, wherein the CNN further comprises a 2D spatial-temporal

feature extracting layer coupled to the swap layer and parallel to the 2D spatial feature extracting layer, and wherein the determining comprises:

- receiving the 2D spatial-temporal feature map from the swap layer;
- enhancing discrepancies of the FOI in the 2D spatial-temporal feature map;
- recognizing the target object according to the enhanced FOI; and
- performing a derivation operation on the recognized target object in the temporal direction, to determine the speed and the acceleration of the target object.

7. The method of claim 6, further comprising:

- training the feature extracting layer independently; and
- training, based on the trained feature extracting layer, the 2D spatial feature extracting layer and the 2D spatial-temporal feature extracting layer separately.

8. The method of claim 3, wherein the evaluating comprises comparing the first feature maps with a binary mask generated from the preset image.

9. A system for tracking a target object in a video, comprising:

- a memory that stores executable components; and
- a processor electrically coupled to the memory to execute the executable components

for:

- extracting, from the video, a 3D feature block containing the target object;
- decomposing the extracted 3D feature block into:
 - a 2D spatial feature map containing spatial information of the target object;
 - and
 - a 2D spatial-temporal feature map containing spatial-temporal information of the target object;
- estimating, in the 2D spatial feature map, a location of the target object;
- determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object;

calibrating the estimated location of the target object according to the determined speed and acceleration; and

tracking the target object in the video according to the calibrated location.

10. The system of claim 9,

wherein the 2D spatial feature map extends in:

a first spatial direction; and

a second spatial direction intersecting with the first spatial direction; and

wherein the 2D spatial-temporal feature map comprises:

a first 2D spatial-temporal feature map extending in the first spatial direction and a temporal direction and comprising components of the speed and the acceleration of the target object in the first spatial direction; and

a second 2D spatial-temporal feature map extending in the second spatial direction and the temporal direction and comprising components of the speed and the acceleration of the target object in the second spatial direction.

11. The system of claim 10, further comprising a CNN including a feature extracting layer, wherein the extracting comprises:

filtering, in the feature extracting layer, each frame of the video to obtain first feature maps;

evaluating overlap degrees and similarity degrees between the first feature maps and a preset image containing a feature of interest (FOI) of the target object; and

selecting, from the first feature maps, a second feature map according to the overlap degrees and the similarity degrees, wherein the second feature map only contains the FOI of the target object; and

combining the selected second feature map over each frame of the video together to construct the 3D feature block.

12. The system of claim 11, wherein the CNN further comprises a swap layer coupled to

the feature extracting layer, and wherein the decomposing comprises:

- receiving the 3D feature block from the feature extracting layer;
- disabling data of the received feature block in the temporal direction to obtain the 2D spatial feature map; and
- disabling data of the received feature block in one of the first spatial direction and the second spatial direction to obtain the 2D spatial-temporal feature map.

13. The system of claim 12, wherein the CNN further comprises a 2D spatial feature extracting layer coupled to the swap layer, and wherein the estimating comprises:

- receiving the 2D spatial feature map from the swap layer;
- enhancing discrepancies of the FOI in the 2D spatial feature map;
- recognizing the target object according to the enhanced FOI; and
- estimating the location of the recognized target object.

14. The system of claim 13, wherein the CNN further comprises a 2D spatial-temporal feature extracting layer coupled to the swap layer and parallel to the 2D spatial feature extracting layer, and wherein the determining comprises:

- receiving the 2D spatial-temporal feature map from the swap layer;
- enhancing discrepancies of the FOI in the 2D spatial-temporal feature map;
- recognizing the target object according to the enhanced FOI; and
- performing a derivation operation on the recognized target object in the temporal direction, to determine the speed and the acceleration of the target object.

15. The system of claim 14, wherein the feature extracting layer is trained independently; and the 2D spatial feature extracting layer and the 2D spatial-temporal feature extracting layer are separately trained based on the trained feature extracting layer.

16. The system of claim 11, wherein the evaluating comprises comparing the first feature maps with a binary mask generated from the preset image.

17. A system for tracking a target object in a video, comprising:

- a feature extractor configured for extracting, from the video, a 3D feature block containing the target object;
- a decomposer configured for decomposing the extracted 3D feature block into:
 - a 2D spatial feature map containing spatial information of the target object; and
 - a 2D spatial-temporal feature map containing spatial-temporal information of the target object;
- a locator configured for estimating, in the 2D spatial feature map, a location of the target object;
- a motion detector configured for determining, in the 2D spatial-temporal feature map, a speed and an acceleration of the target object;
- a calibrator configured for calibrating the estimated location of the target object according to the determined speed and acceleration; and
- a tracker configured for tracking the target object in the video according to the calibrated location.

18. The system of claim 17,

wherein the 2D spatial feature map extends in:

- a first spatial direction; and
- a second spatial direction intersecting with the first spatial direction; and

wherein the 2D spatial-temporal feature map comprises:

- a first 2D spatial-temporal feature map extending in the first spatial direction and a temporal direction and comprising components of the speed and the acceleration of the target object in the first spatial direction; and
- a second 2D spatial-temporal feature map extending in the second spatial direction and the temporal direction and comprising components of the speed and the acceleration of the target object in the second spatial direction.

19. The system of claim 18, further comprising a CNN including a feature extracting layer, wherein the feature extractor is implemented in the feature extracting layer and configured for:

filtering, in the feature extracting layer, each frame of the video to obtain first feature maps;

evaluating overlap degrees and similarity degrees between the first feature maps and a preset image containing a feature of interest (FOI) of the target object; and

selecting, from the first feature maps, a second feature map according to the overlap degrees and the similarity degrees, wherein the second feature map only contains the FOI of the target object; and

combining the selected second feature map over each frame of the video together to construct the 3D feature block.

20. The system of claim 19, wherein the CNN further comprises a swap layer coupled to the feature extracting layer, and wherein the decomposer is implemented in the swap layer and configured for:

receiving the 3D feature block from the feature extracting layer;

disabling data of the received feature block in the temporal direction to obtain the 2D spatial feature map; and

disabling data of the received feature block in one of the first spatial direction and the second spatial direction to obtain the 2D spatial-temporal feature map.

21. The system of claim 20, wherein the CNN further comprises a 2D spatial feature extracting layer coupled to the swap layer, and wherein the locator is implemented in the 2D spatial feature extracting layer and configured for:

receiving the 2D spatial feature map from the swap layer;

enhancing discrepancies of the FOI in the 2D spatial feature map;

recognizing the target object according to the enhanced FOI; and

estimating the location of the recognized target object.

22. The system of claim 21, wherein the CNN further comprises a 2D spatial-temporal feature extracting layer coupled to the swap layer and parallel to the 2D spatial feature extracting layer, and wherein the motion detector is implemented in the 2D spatial-temporal feature extracting layer and configured for:

- receiving the 2D spatial-temporal feature map from the swap layer;
- enhancing discrepancies of the FOI in the 2D spatial-temporal feature map;
- recognizing the target object according to the enhanced FOI; and
- performing a derivation operation on the recognized target object in the temporal direction, to determine the speed and the acceleration of the target object.

23. The system of claim 22, wherein the feature extracting layer is trained independently; and the 2D spatial feature extracting layer and the 2D spatial-temporal feature extracting layer are separately trained based on the trained feature extracting layer.

24. The system of claim 19, wherein the evaluating is implemented in a comparator for comparing the first feature maps with a binary mask generated from the preset image.

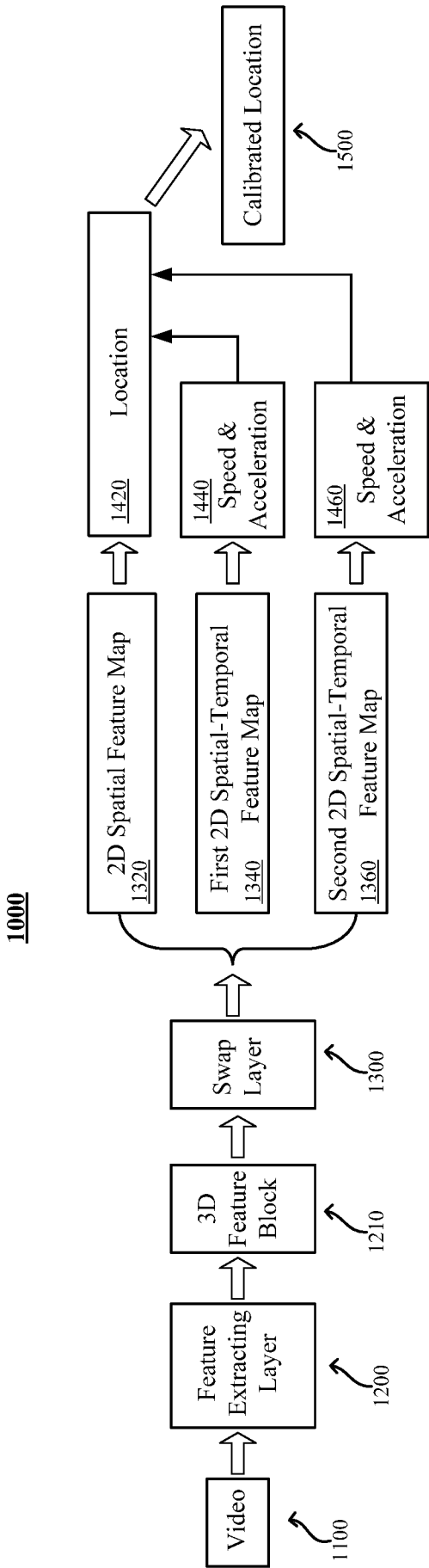
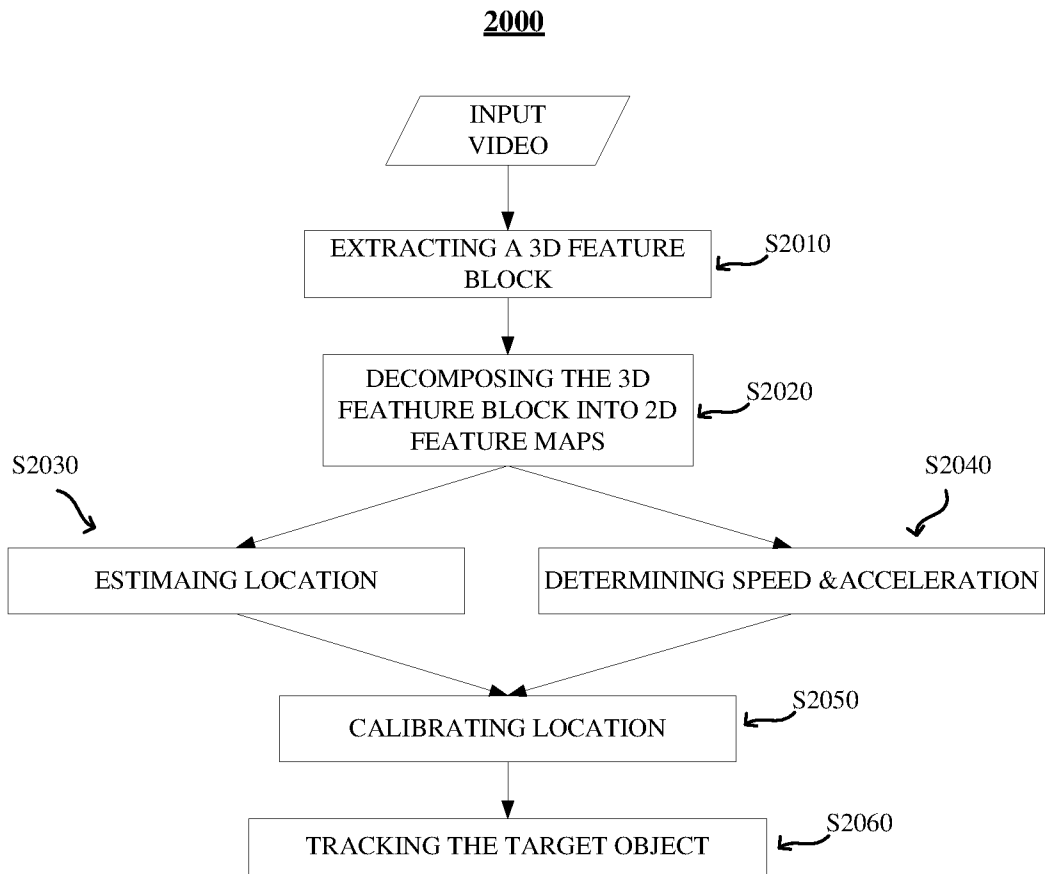


Fig. 1



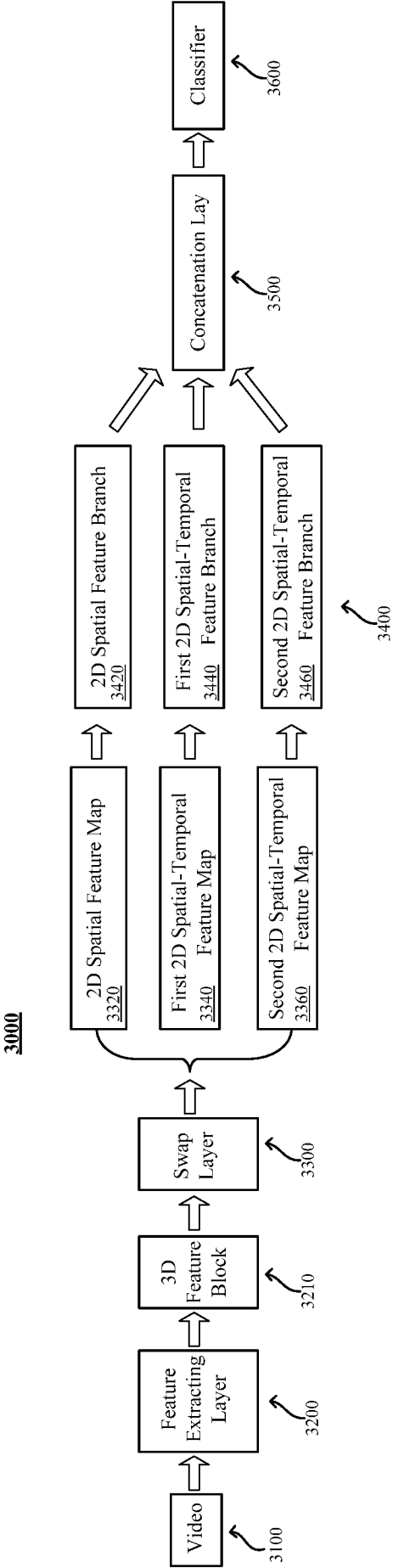


Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/078982

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00(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, H04N

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)

CNKI, CNPAT, WPI, EPODOC: object, target, track, trace, trail, detect, recognise, spatial, spatio, temporal, XT, YT, XY, XYT, speed, acceleration, location, calibrate, adjust, rectify, modify, correct, revise, CNN, convolutional neural network.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2010034423 A1 (ZHAO, TAO ET AL.) 11 February 2010 (2010-02-11) the abstract, description, paragraphs [0020]-[0035], and figures 1A-1B	1-24
A	CN 101827224 A (HOHAI UNIVERSITY) 08 September 2010 (2010-09-08) the whole document	1-24
A	US 2011182469 A1 (NEC LABORATORIES AMERICA, INC.) 28 July 2011 (2011-07-28) the whole document	1-24
A	US 2010201871 A1 (ZHANG, DONG-QING ET AL.) 12 August 2010 (2010-08-12) the whole document	1-24
A	RICQUEBOURG, Yann et al. "Real-Time Tracking of Moving Persons by Exploiting Spatio-Temporal Image Slices" <i>IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE</i> , Vol. 22, No. 8, 31 August 2000 (2000-08-31), the whole document	1-24

☒ 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

20 December 2016

Date of mailing of the international search report

30 December 2016

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088
China

Authorized officer

JIN,Xiaocong

Facsimile No. (86-10)62019451

Telephone No. (86-10)62413230

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/078982**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	FAN, Jialue et al. "Human Tracking Using Convolutional Neural Networks" <i>IEEE TRANSACTIONS ON NEURAL NETWORKS</i> , Vol. 21, No. 10, 31 October 2010 (2010-10-31), the whole document	1-24
A	Ji, Shuiwang et al. "3D Convolutional Neural Networks for Human Action Recognition" <i>IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE</i> , Vol. 35, No. 1, 31 January 2013 (2013-01-31), the whole document	1-24

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/078982

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US	2010034423	A1	11 February 2010	US 2012207353 A1	16 August 2012
CN	101827224	A	08 September 2010	None	
US	2011182469	A1	28 July 2011	None	
US	2010201871	A1	12 August 2010	None	