



(51) International Patent Classification:

G06F 17/16 (2006.01) G06N 3/08 (2006.01)  
G06K 9/00 (2022.01) G06T 11/00 (2006.01)  
G06K 9/46 (2006.01) G06T 11/60 (2006.01)

(21) International Application Number:

PCT/US2022/022112

(22) International Filing Date:

28 March 2022 (28.03.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/170,230 02 April 2021 (02.04.2021) US

(71) Applicant: CARNEGIE MELLON UNIVERSITY

[US/US]; 5000 Forbes Avenue, Pittsburgh, PA 15213 (US).

(72) Inventors; and

(71) Applicants: SAVVIDES, Marios [US/US]; 506 Eagle Court, Wexford, PA 15090 (US). HU, Kai [US/US]; 101 Shady Ave., Apt. D607, Pittsburgh, PA 15206

(US). ZHENG, Yutong [US/US]; 5000 Forbes Avenue, Pittsburgh, PA 15213 (US). AHMED, Uzair [US/US]; 5000 Forbes Avenue, Pittsburgh, PA 15213 (US). NALLAMOTHU, Sreena [US/US]; 5000 Forbes Avenue, Pittsburgh, PA 15213 (US).

(74) Agent: CARLETON, Dennis, M.; 2601 Weston Parkway, Suite 103, Cary, NC 27513 (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, IT, JM, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: SYSTEM AND METHOD FOR POSE TOLERANT FEATURE EXTRACTION USING GENERATED POSE-ALTERED IMAGES

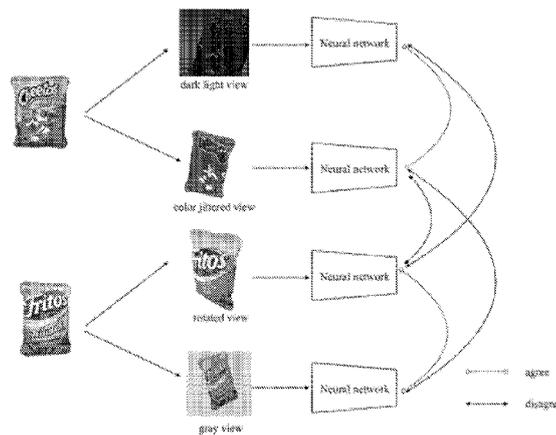


FIG. 1

(57) Abstract: Disclosed herein is a system and method for augmenting data by generating a plurality of pose-altered images of an item from one or more 2D images of the item and using the augmented data to train a feature extractor. In other aspects of the invention, the trained feature extractor is used to enroll features extracted from images of new products in a library database of known products or to identify images of unknown products by matching features of an image of the unknown product with features stored in the library database.



**(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))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

## **System and Method for Pose Tolerant Feature Extraction Using Generated Pose-Altered Images**

### **Related Applications**

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 63/170,230, filed April 2, 2021, the contents of which are incorporated herein in their entirety.

### **Background**

[0002] In a retail setting, it is desirable to be able to use computer vision methods to detect and identify products on a retail shelf to aid in management of the retail establishment. For example, computer vision may be used to detect and identify products for various tasks, such as tracking product inventory, determining out-of-stock products and determining misplaced products. Product detection is one of the fastest-moving areas and plays a fundamental role in many retail applications such as product recognition, planogram compliance, out-of-stock management, and check-out free shopping.

[0003] To this end, numerous computer vision methods have been developed and many real-world applications based on those computer vision methods perform at a satisfactory level. Currently, various visual sensors (e.g., fixed cameras, robots, drones, and mobile phones) have been deployed in retail stores, enabling the application of advanced technologies to ease shopping and store management tasks.

[0004] Feature extractors typically extract features that are used in downstream tasks, such as classification. However, products can be of arbitrary poses in a real-world retail scene, especially when the image is taken by a camera not facing straight towards the shelf. For example, images could show products at arbitrary angles, rotated, crumpled (e.g., in the case of products packaged in bags, such as potato chips), color jittered, over-exposed, under-exposed, etc. Because of these difficulties, features extracted from images collected in a retail setting may not be able to be matched to features extracted from a pristine image of the product, such as an image of the product provided by the manufacturer. Therefore, the features extracted from these images may not be accurate for the downstream tasks.

### **Summary**

[0005] For building a system that is robust for product views, a conventional strategy is to train the artificial intelligence models based on the downstream task with overly sufficient data spanning across the different views of the objects. Collecting and accumulating qualified and representative datasets is not easy in the real world due to various time, labor and financial constraints. Data becomes the real bottleneck in the learning capacity of many of the machine learning and artificial intelligence models.

[0006] Described herein is a system and method for data augmentation strategies to generate machine learning training data that is crucial in training efficient models when the supply of training data is limited.

[0007] The disclosed system and method uses a neural network to learn the feature embeddings of objects in different object pose views (2D rotation, 3D rotation, light, resolution, etc.) to learn pose-invariant feature embeddings, neural representations that are invariant to both the 2D and 3D orientation of an object. The method trains the neural network using data augmented by a plurality of 2D views of the product.

[0008] In one embodiment, the plurality of 2D views are generated by first generating a 3D model of the product from a 2D image of the product, and then generating novel, synthetic 2D images of how the product looks at varying poses. Preferably, the 2D image from which the synthetic, pose-altered images are generated is a pristine image of the product supplied by the manufacturer, but any image may be used. In other embodiments, different methods of generating the synthetic, pose-altered 2D views are disclosed.

[0009] In a second aspect of the invention, features extracted from the varying views are enrolled in a database and used for product classification and, in a third embodiment, images of an unknown testing product are matched against a library database to attempt identification of the unknown product.

**Brief Description of the Drawings**

[0010] By way of example, a specific exemplary embodiment of the disclosed system and method will now be described, with reference to the accompanying drawings, in which:

[0011] **FIG. 1** is an illustration showing training of a feature extractor using augmented views as disclosed herein.

[0012] **FIG. 2** is an illustration of a first aspect of the invention for generating the augmented views from a single image of the product.

[0013] **FIG. 3** is an illustration of a second aspect of the invention for generating the augmented views from a single image of the product.

[0014] **FIG. 4** is an illustration of a third aspect of the invention for generating the augmented views from a single image of the product.

[0015] **FIG. 5** is an illustration of a fourth aspect of the invention for generating the augmented views from a single image of the product.

[0016] **FIG. 6** is an illustration of the process of enrolling extracted feature in a product database library.

[0017] **FIG. 7** is a flowchart depicting the training process of the feature extractor.

[0018] **FIG. 8** is a flowchart depicting the process of enrolling a new item in the library database.

[0019] **FIG. 9** is a flowchart depicting the process of matching an unknown item with items enrolled in the library database.

**Detailed Description**

[0020] In the retail industry, new products are manufactured and brought to market on a regular basis. Additionally, existing products are always subject to a change in their packaging due to new designs or seasonal packaging throughout the year. To maintain a current library of products, it is important to be able to enroll and re-enroll product images in a timely manner.

[0021] To create an over-complete library of products with different variations in pose, the disclosed system and method first creates multiple 2D pose-altered images of the product. In a preferred embodiment, the multiple 2D pose-altered images are created by first creating a 3D model of the product from a high resolution 2D product image, which is often supplied by the manufacturer of the product. The 3D models can then be rotated along different axes, thereby providing access to different views of the same product from different view-points. Other method of generating the 2D pose-altered images of the product are also disclosed herein

[0022] The method then extracts features from the generated pose-altered product images and adds these features to a library. This reduces the domain gap between the images in the library and the product images coming from store cameras, thus making the matching process invariant to the pose of the product to be identified. This results in a robust matching algorithm to recognize the product.

[0023] This approach creates a library of features representing products from different viewpoints. Each of the pose-altered images (as well as the original image of the product) are then passed through a feature extractor to obtain features of the images, which are then stored in a library database.

[0024] To train the feature extraction method for classification with a limited dataset access of products at different angles, the invention employs a training method that uses generation of 3D models of products from 2D images. These 3D models can be used to synthesize novel 2D images of the product from various viewpoints that can be used as a training dataset for building a classifier or a feature representation/extraction method (e.g., a feature extractor). The described method is key to building any type of pose-tolerant feature extraction method.

[0025] In one embodiment, to build a robust feature extractor for product identification regardless of data type, a self-supervised method is applied to augment the database for training. The method is a mixture of data augmentation strategies and multiple loss functions. A major performance improvement is achieved in comparison to traditional classification methods. The method is independent of model architecture and other existing components for classification tasks.

[0026] To achieve this, the method specifies a metric learning that maximizes the agreement/similarity between augmented images of the same product and minimizes that between different products in the feature space. After a good

feature extractor is learned, many downstream tasks can be greatly improved using features extracted from the products, as shown in **FIG. 1**. In addition, ring loss, as described in U.S. Published Patent Application No. 2021/0034984, may be applied along with self-supervised learning to improve the model's capability to do classification for associated tasks.

[0027] For product identification, there is typically an onboarding product approach.

One such product identification generates a 3D model. The method and process of feature extraction using 3D models can be applied for enrolling an image to onboard a new product using a 3D model generated from 2D images or other descriptions of the product with or without pose variations.

[0028] Several aspects of the invention specify different ways to generate the 3D model or to otherwise generate augmented 2D images of the product having alternate poses.

[0029] In a first aspect of the invention, shown in **FIG. 2**, product views are generated using 3D model generation. In this approach, novel product views are generating a 3D model of the product , perturbing the 3D model in a plurality of different ways and capturing 2D views from the perturbations of the 3D model. The 3D model can be perturbed by rotation of the product along any or all of the 3 axes for any combination of degrees of rotation, thus generating product views from different viewpoints.

[0030] In a second aspect of the invention, shown in **FIG. 3**, product views are generated using a 3D planar warp. In this method, explicit creation of the 3D

model is avoided. The 2D image is fixed onto a plane with a known depth.

Because the depth of the plane is known, the plane can be rotated along the 3 different axes to generate product images from different viewpoints.

[0031] In a third aspect of the invention, shown in **FIG. 4**, product views are generated by 2D projective transforms. Planar homography is a powerful tool that can be leveraged to fit images to arbitrary poses as long as all of the corresponding points maintain the same depth. In this approach, novel views are generated by fitting the corner endpoints of the image to different configurations, thus simulating different views.

[0032] In a fourth aspect of the invention, shown in **FIG. 5**, a machine learning-based data generation method is used. Generative modeling is a well-known technique in machine learning where a distribution of the input samples is learned from the labels, where sampling the distribution leads to the generation of novel data samples. Various methods include, but are not limited to, the use of autoencoders and adversarial learning based methods, spatial transformer networks where novel product views may be generated from various input reference images.

[0033] For any of the aspects of the invention for generation of the pose-altered views of the product, in the case wherein products have more than one side visible in the 2D image, an additional optional step may be added to first pose-correct the image such that only the front facing side of the product is visible. The frontal pose correction to get the front face of the product may

or may not be done based on the application for which the data is being augmented.

[0034] As would be realized by one of skill in the art, any method of generating multi-pose images of an item from a single 2D image may be used to augment the training dataset, as well as to provide additional views for feature extraction, wherein the extracted features are enrolled in the library for later matching with features extracted from live images of the products.

[0035] The synthesized, pose-altered, novel views of a product generated by data augmentation approaches such as any of the methods in the four aspects described herein, or by any other method, can be used for enrollment of the product in a product identification system, as shown in **FIG. 6**. Augmented image data can be used to extract features to be enrolled in a library database for further matching or training. The augmented image data can be sent to a database for 1:1 or 1:N verifications using an image classifier.

[0036] This method and process applies to any machine learning and artificial intelligence system using only a single reference image of a product by generating various product views for matching uses. That reference image can be a retail product image or a single captured image of a product from any capture source.

[0037] In some aspects of the invention, the features are extracted from all or partially synthesized or augmented views and are matched with features extracted from an image of an unknown testing product, which may have

been captured, for example, by a robotic inventory system of a retail establishment. This matching provides a very efficient way to enroll products. It also provides the matching at any angle based on the reference image when only a single pose image of the product is available.

[0038] In another aspect of the invention, if more than one reference image is available, then more views can be generated and more image variations of the product become available for training, learning, and detection. For example, augmented 3D models may generate front and back views of a product for use by the learning and detection models. This provides the system a full view of the product. Thus, using the augmented 3D models of the products in an artificial intelligence model for object detection, or the techniques and approaches outlined above, the system generates multiple images from various viewpoints to simulate the camera settings and reduce the domain gap between the training and testing scenarios.

[0039] **FIG. 7** is a flowchart showing the process of training the feature extractor. 2D images of various items are gathered and, at step 702, synthetic images showing multiple poses of the items in the 2D images are generated from each of the 2D images. In preferred embodiments of the invention one of the methods described herein for generating the synthetic, pose-altered images may be used. In alternate embodiments, any other method of generating the synthetic, pose-altered images may be used. The original 2D images, along with the synthetic pose-altered images comprise the training dataset for the

feature extractor. At 704, the feature extractor is trained on the training dataset, resulting in a feature extractor trained to extract features from images showing multiple poses of a single item, which may be used for product enrollment or for product identification, as described with respect to **FIG. 8** and **FIG. 9**.

[0040] **FIG. 8** is a flowchart showing the process of enrolling a new product in the library database. A 2D image of the new item is used to generate multiple synthetic, pose-altered images of the new item at step 802, resulting in a set of images of the new item with multiple poses (including the original 2D image of the item). The set of images of the new item is then input to the feature extractor at step 804. The feature extractor extracts features of the images and, at step 806, the features of the images are enrolled in the library database for use in product identification, as described with respect to **FIG. 9**.

[0041] **FIG. 9** is a process of matching an image of an unknown item with a known image. The 2D image of the unknown item may have any pose. At 902, the image of the unknown item is input to the feature extractor. The feature extractor outputs features of the image of the unknown item and, at step 904, an attempt is made to match the features of the image of the unknown item with features of images of items in the library database, including the pose-altered images of the items in the library database. At 906, if there is a match, the image is positively identified. At 906, if no match is found, the images not identified. The unidentified image may then optionally be

enrolled in the library database by the process shown in **FIG. 8**. The matching may be performed by a trained classifier.

[0042] As would be realized by one of skill in the art, the disclosed method described herein can be implemented by a system comprising a processor and memory, storing software that, when executed by the processor, performs the functions comprising the method.

[0043] As would further be realized by one of skill in the art, many variations on implementations discussed herein which fall within the scope of the invention are possible. Moreover, it is to be understood that the features of the various embodiments described herein were not mutually exclusive and can exist in various combinations and permutations, even if such combinations or permutations were not made express herein, without departing from the spirit and scope of the invention. Accordingly, the method and apparatus disclosed herein are not to be taken as limitations on the invention but as an illustration thereof. The scope of the invention is defined by the claims which follow.

**Claims**

1. A method comprising:
  - generating a plurality of pose-altered images of a plurality of items based on one or more 2D images of each of the plurality of items; and
  - training a feature extractor to extract features of the synthetic pose-altered images.
  
2. The method of claim 1 wherein the pose-altered images include views of the items which are modified from the one or more 2D images of the items, the modifying including presenting the items at arbitrary angles, rotating the items, translating the items, crumpling the items, modifying the colors of the items, over-exposing the items and under-exposing the items.
  
3. The method of claim 2 further comprising:
  - generating a plurality of pose-altered images of a new item based on one or more 2D images of the new item;
  - inputting the pose-altered images of the new item to the feature extractor to obtain features of the pose-altered images; and
  - enrolling the extracted features in a library database.

4. The method of claim 3 wherein the one or more 2D images of the new item are also input to the feature extractor.
  
5. The method of claim 2 further comprising:
  - identifying an unknown item in a testing image by:
    - exposing the testing image of the unknown item to the feature extractor to obtain features of the testing image; and
    - matching the features extracted from the testing image with features enrolled in the library database to determine a match to identify the unknown item.
  
6. The method of claim 5 wherein the matching is performed by a trained classifier.
  
7. The method of claim 2 wherein the step of generating a plurality of pose-altered images comprises:
  - generating a 3D model of the item from the one or more 2D images of the item;
  - generating images showing different viewpoints of the item by rotating the 3D model along one or more axes.

8. The method of claim 2 wherein the step of generating a plurality of pose-altered images comprises:
  - fixing the one or more 2D images of the item to a plane with a known depth; and
  - generating images showing different viewpoints of the item by rotating the 2D images along one or more of the axes of the plane.
  
9. The method of claim 2 wherein the step of generating a plurality of pose-altered images comprises:
  - using planar homography by fitting the corner endpoints of the one or more 2D images to different configurations to simulate different views of the item.
  
10. The method of claim 2 wherein the step of generating a plurality of pose-altered images comprises:
  - training a machine learning model to generate the pose-altered images; and
  - inputting the one or more 2D images of the item to the machine learning model.
  
11. The method of claim 2 further comprising:

pose correcting the one or more 2D images of the item to eliminate portions of the image containing views of sides of the item other than the frontal side.

12. A system comprising:

a processor; and

software that, when executed by the processor, causes the system to:

generate a plurality of pose-altered images of a plurality of items based on one or more 2D images of each of the plurality of items; and

train a feature extractor to extract features of the pose-altered images.

13. The system of claim 12 wherein the pose-altered images include views of the items which are modified from the one or more 2D images of the items, the modifying including presenting the item at arbitrary angles, rotating the item, translating the item, crumpling the item, modifying the colors of the item, over-exposing the item and under-exposing the item.

14. The system of claim 13 the software further causing the system to:

generate a plurality of pose-altered images of a new item based on one or more 2D images of the new item;

input the one or more pose-altered images of the new item to the feature extractor to obtain features of the pose-altered images; and enroll the extracted features in a library database.

15. The system of claim 14, the software further causing the system to:

input the one or more 2D images of the new item are to the feature extractor.

16. The system of claim 13, the software identifying an unknown item in a testing image by causing the system to:

expose the testing image of the unknown item to the feature extractor to obtain features of the testing image; and match the features extracted from the testing image with features enrolled in the library database to determine a match to identify the unknown item.

17. The system of claim 16 wherein the matching is performed by a trained classifier.

18. The system of claim 13 wherein the software causes the generation of the plurality of pose-altered images by causing the system to:

generate a 3D model of the item from the one or more 2D images of the item; and

generate images showing different viewpoints of the item by rotating the 3D model along one or more of the axes.

19. The system of claim 13 wherein the software causes the generation of the plurality of pose-altered images by causing the system to:

fix the one or more 2D images of the item to a plane with a known depth; and

generate images showing different viewpoints of the item by rotating the 2D images along one or more of the axes of the plane.

20. The system of claim 13 wherein the software causes the generation of the plurality of pose-altered images by causing the system to:

use planar homography by fitting the corner endpoints of the one or more 2D images to different configurations to simulate different views of the item.

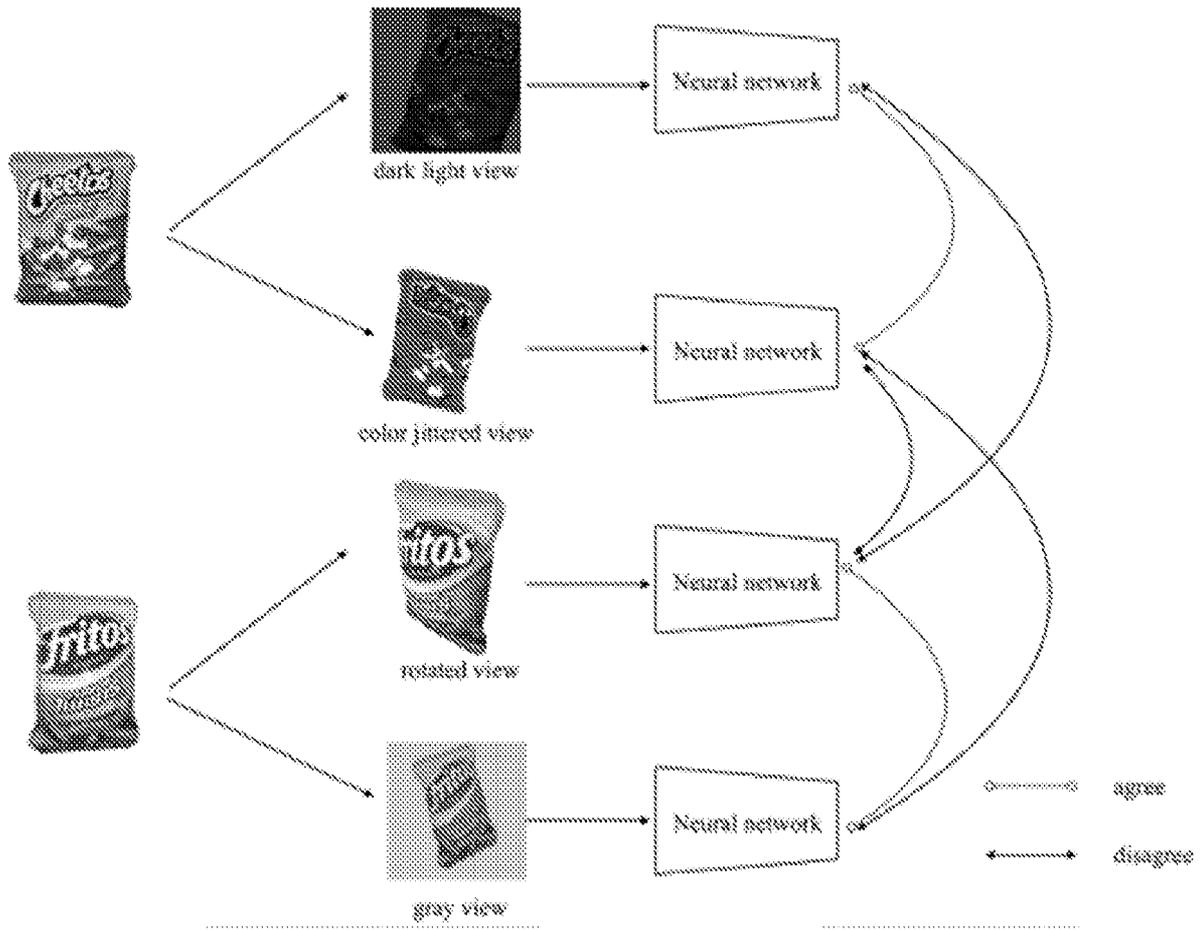
21. The system of claim 13 wherein software causes the generation of the plurality of pose-altered images by causing the system to:

train a machine learning model to generate the pose-altered images;  
and

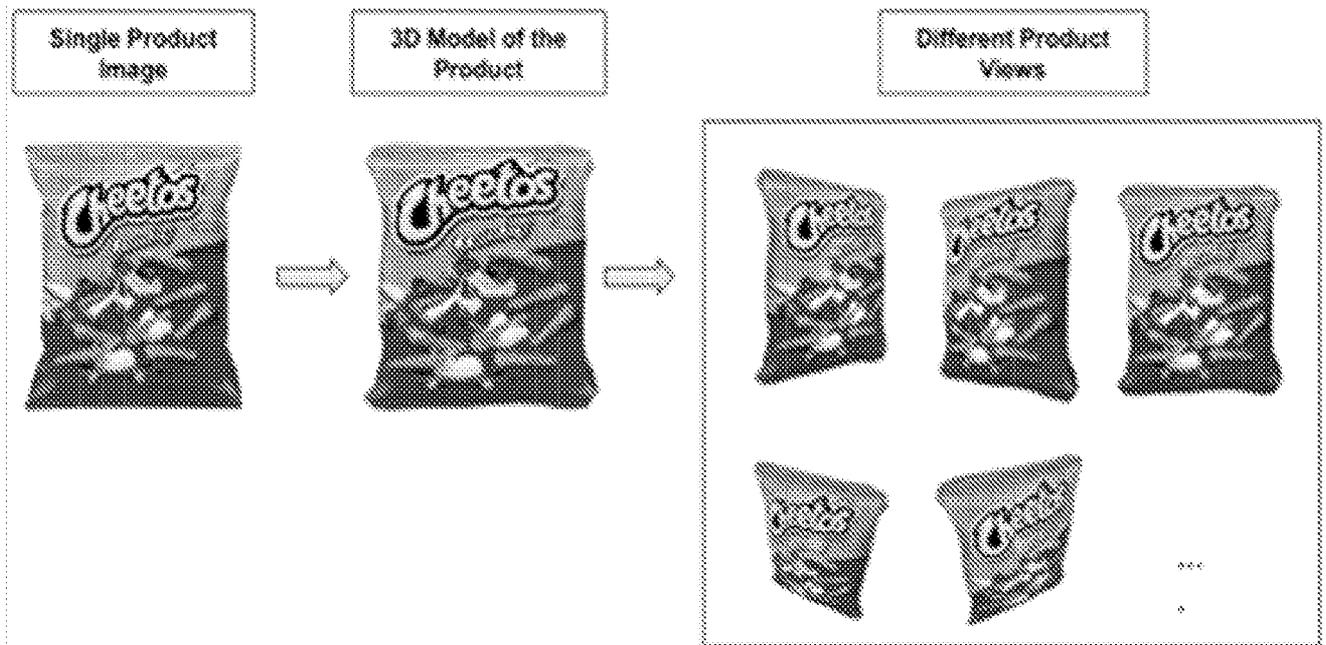
input the one or more 2D images of the item to the machine learning model.

22. The system of claim 13 wherein the software further causes the system to:

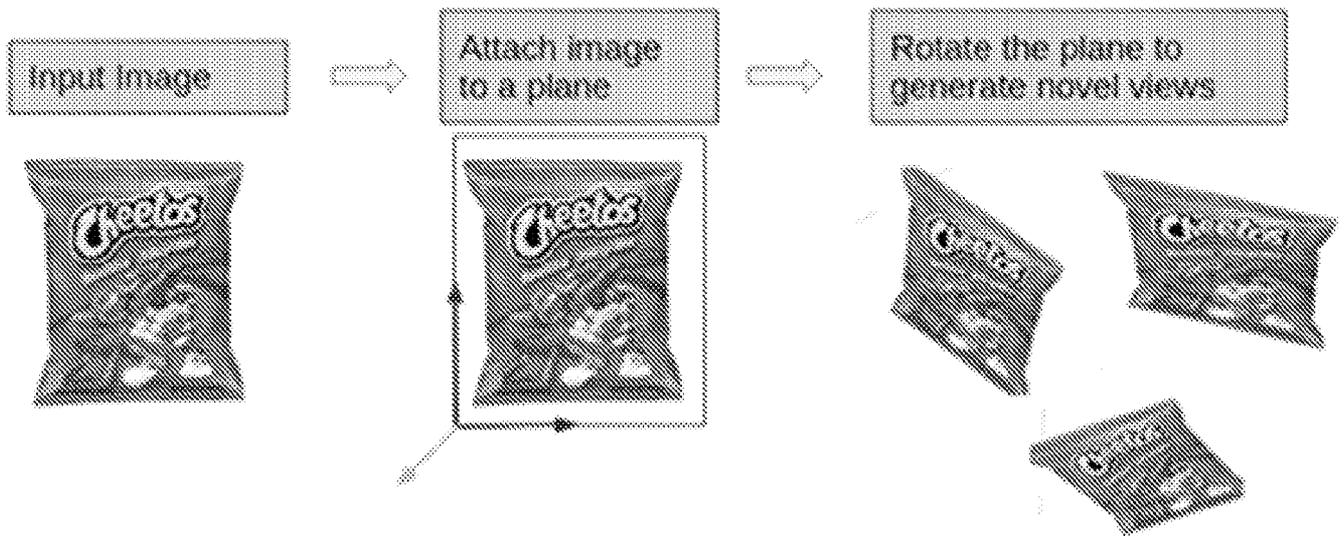
pose correct the one or more 2D images of the item to eliminate portions of the one or more images containing views of sides of the item other than the frontal side.



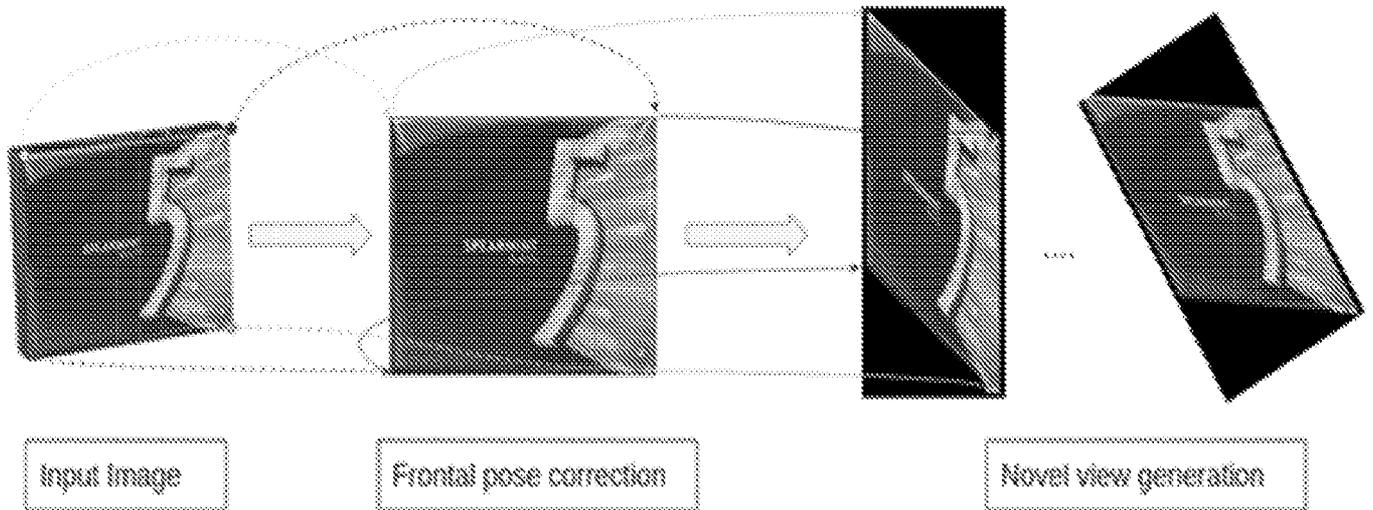
**FIG. 1**



**FIG. 2**



**FIG. 3**



**FIG. 4**

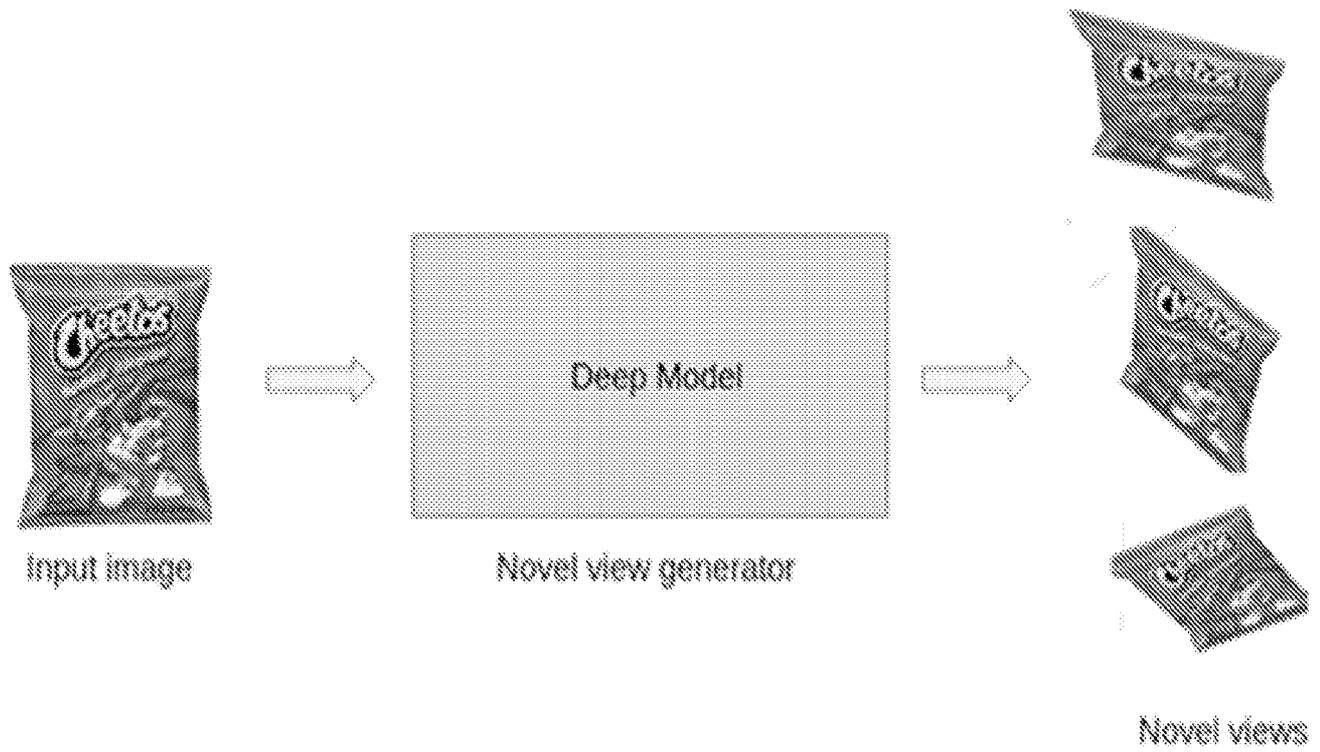
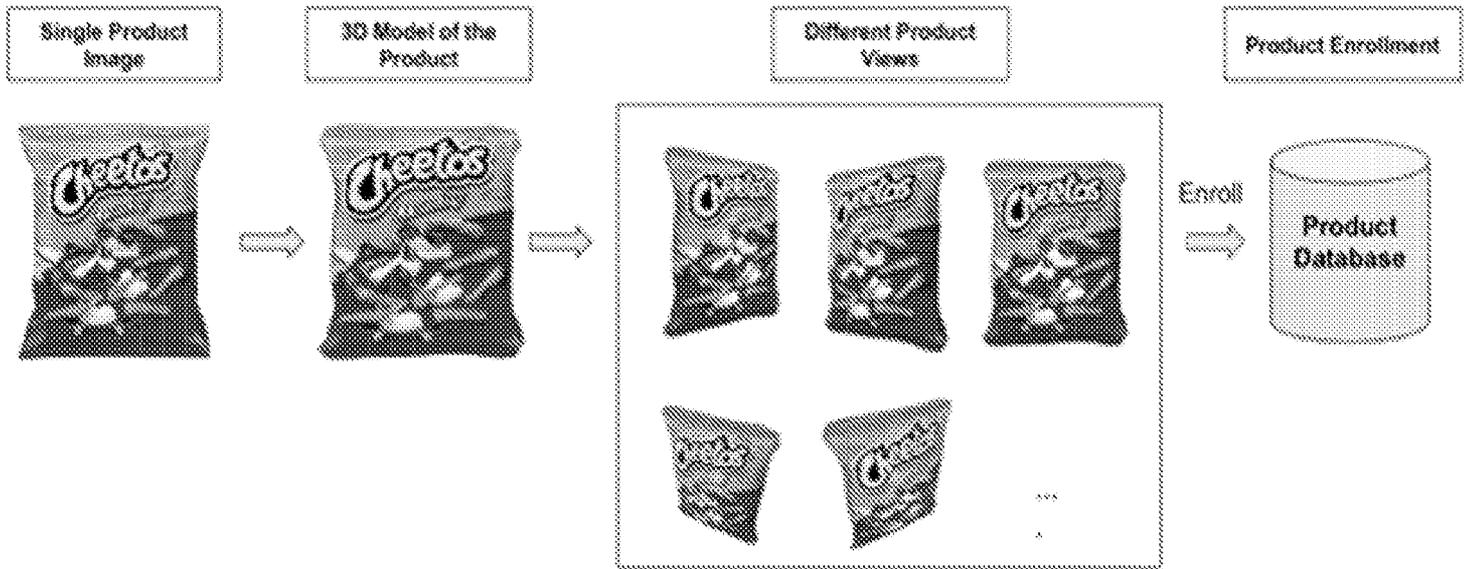
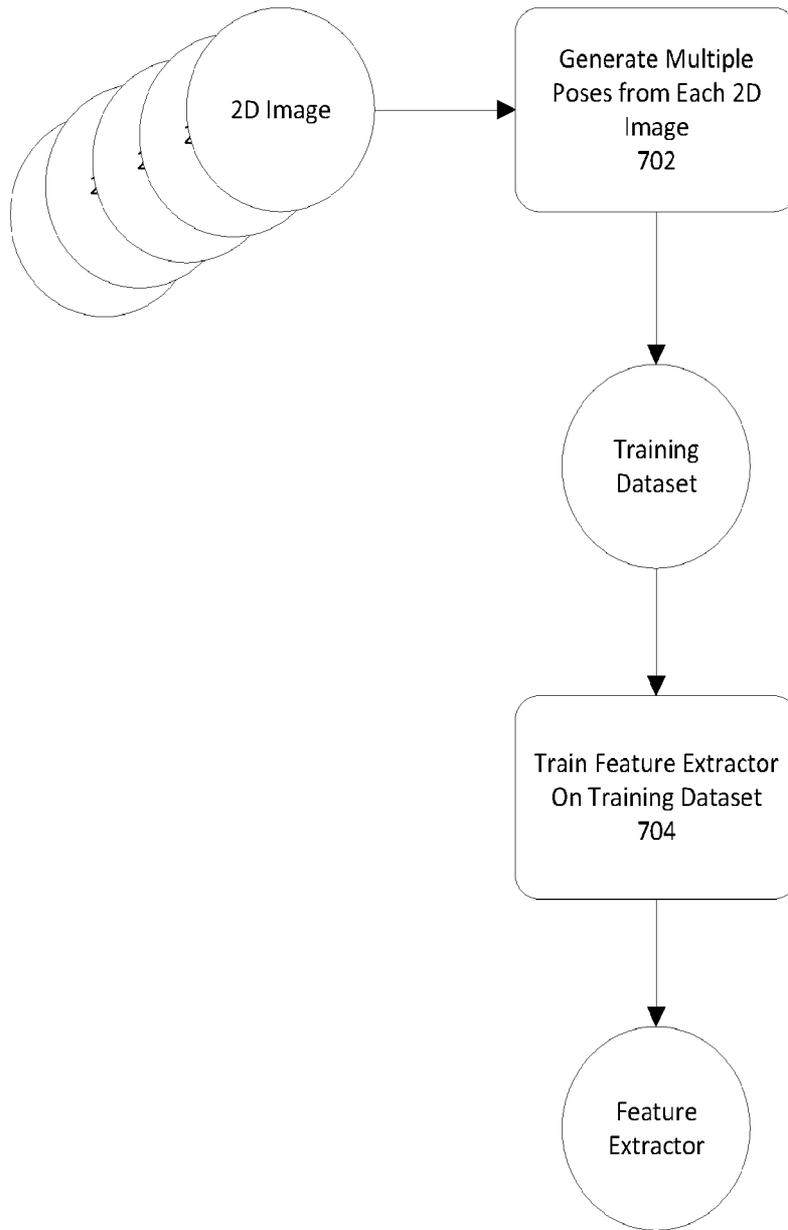


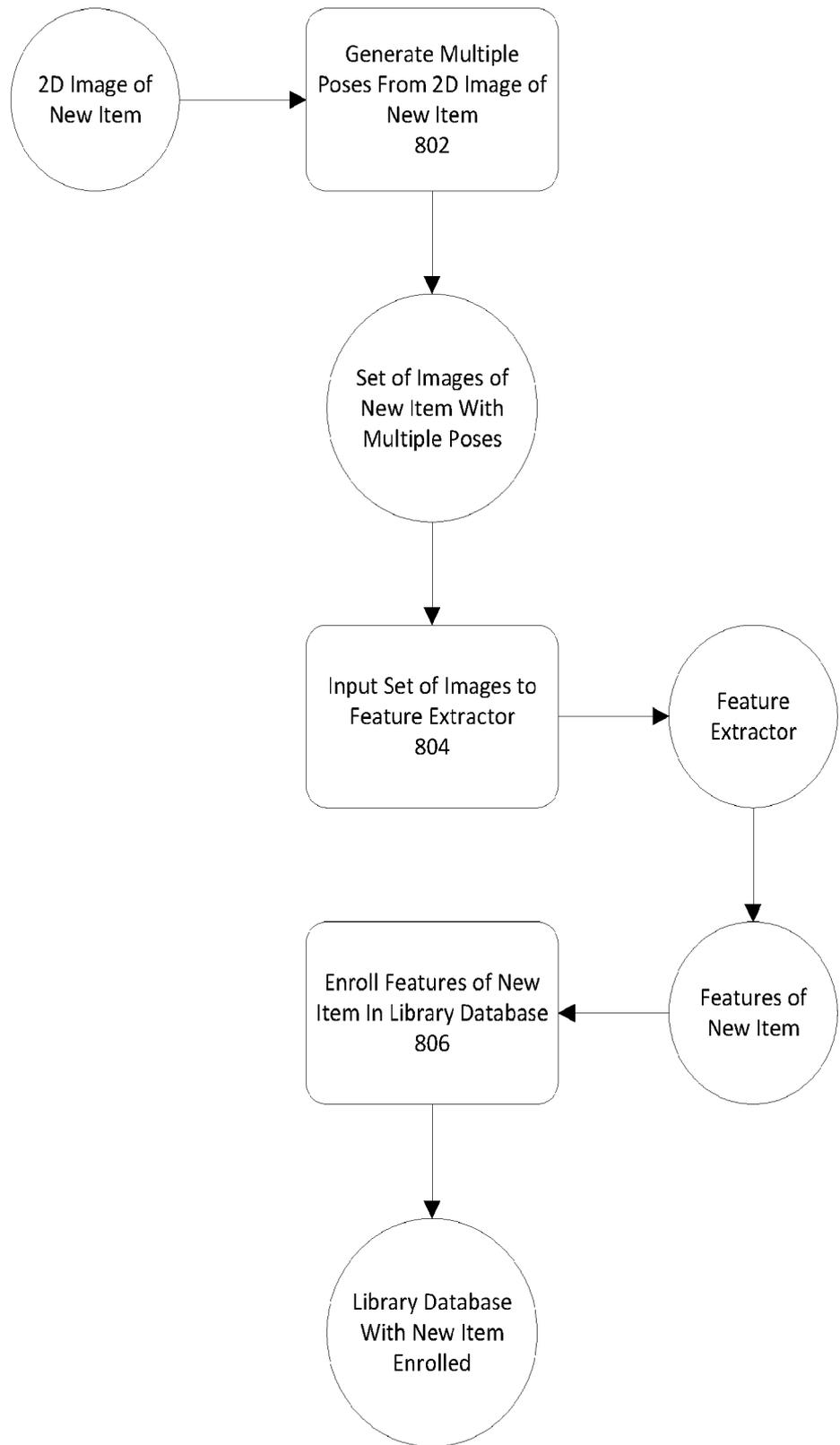
FIG. 5



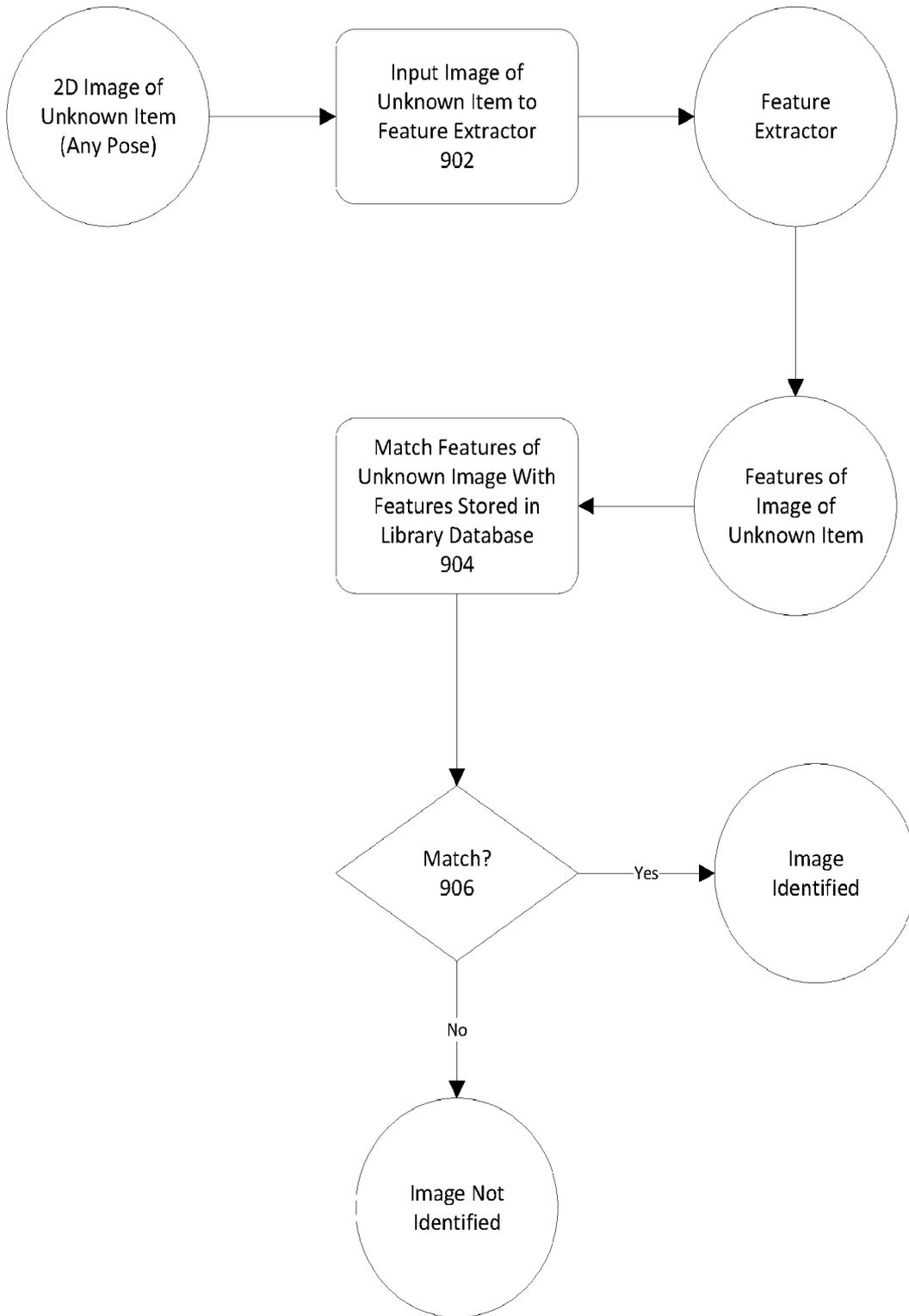
**FIG. 6**



**FIG. 7**



**FIG. 8**



**FIG. 9**

INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US2022/022112

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 17/16; G06K 9/00; G06K 9/46; G06N 3/08; G06T 11/00; G06T 11/60 (2022.01)  
CPC - G06F 16/5854; G06F 16/5838; G06F 17/16; G06F 2207/4824; G06T 11/001; G06T 11/60; G06T 2207/30201 (2022.05)

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)  
see Search History document

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- A	US 2006/0067573 A1 (PARR et al) 30 March 2006 (30.03.2006) entire document	1, 12 --- 2-11, 13-22
A	BROWNLEE. "How to Configure Image Data Augmentation in Keras." Machine Learning Mastery. 12 April 2019 (12.04.2019) Retrieved on 01 July 2022 (01.07.2022) from < <a href="https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/">https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/</a> > entire document	1-22
A	US 2016/0148074 A1 (CAPTRICITY, INC.) 26 May 2016 (26.05.2016) entire document	1-22
A	US 2021/0034984 A1 (CARNEGIE MELLON UNIVERSITY) 04 February 2021 (04.02.2021) entire document	1-22
A	US 2019/0122404 A1 (HOLITION LIMITED) 25 April 2019 (25.04.2019) entire document	1-22

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

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"D" document cited by the applicant in the international application	"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
"E" earlier application or patent but published on or after the international filing date	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 03 July 2022	Date of mailing of the international search report <b>JUL 25 2022</b>
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Taina Matos Telephone No. PCT Helpdesk: 571-272-4300