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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 June 2020 (11.06.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/118093 A1

(51) International Patent Classification:
G06T 5/20 (2006.01) *G06T 7/20* (2017.01)
G06T 7/00 (2017.01) *H04N 5/232* (2006.01)

(21) International Application Number:
PCT/US2019/064755

(22) International Filing Date:
05 December 2019 (05.12.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/775,725 05 December 2018 (05.12.2018) US

(71) Applicant: **DUKE UNIVERSITY** [US/US]; 2812 Erwin Road, Suite 306, Durham, North Carolina 27705 (US).

(72) Inventors: **BRADY, David Jones**; 8004 Morrell Lane, Durham, North Carolina 27713 (US). **WANG, Chengyu**; 2812 Erwin Road, Suite 306, Durham, North Carolina 27705 (US).

(74) Agent: **BROSEMER, Jeffery J.**; Kaplan Breyer Schwarz, LLP, 90 Matawan Road, Suite 201, Matawan, New Jersey 07747 (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,

(54) Title: NEURAL NETWORK FOCUSING FOR IMAGING SYSTEMS

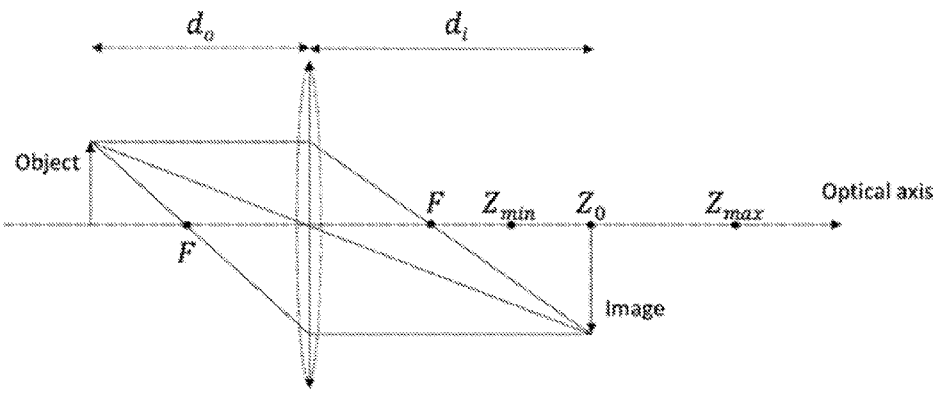


FIG. 1

(57) Abstract: Aspects of the present disclosure describe systems, methods and structures providing neural network focusing for imaging systems that advantageously may identify regions or targets of interest within a scene and selectively focus on such target(s); process subsampled or compressively sampled data directly thereby substantially reducing computational requirements as compared with prior-art contemporary systems, methods, and structures; and may directly evaluate image quality on image data - thereby eliminating need for focus specific hardware.

WO 2020/118093 A1

[Continued on next page]

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))*

NEURAL NETWORK FOCUSING FOR IMAGING SYSTEMS

TECHNICAL FIELD

[0001] This disclosure relates generally to systems, methods, structures, and techniques pertaining to neural network focusing for imaging systems resulting in enhanced image quality.

BACKGROUND

[0002] As is known in the imaging arts, focus – and the ability of an imaging system to quickly and accurately focus - is a primary determinant of image quality. Typically, focus is determined and subsequently controlled by optimizing image quality metrics including contrast and/or weighted Laplacian measures. Such metrics are described, for example by Yao, Y., B. Abidi, N. Doggaz and M. Abidi in an article entitled "Evaluation of Sharpness Measures and Search Algorithms for the Auto-Focusing of High-Magnification Images," that appeared in Defense and Security Symposium, SPIE, (1960). Strategies for optimizing such metrics are described by Jie, H., Z. Rongzhen and H. Zhiliang in an article entitled "Modified Fast Climbing Search Auto-Focus Algorithm with Adaptive Step Size Searching Technique for Digital Camera" that appeared in IEEE Transactions on Consumer Electronics 49(2): 257-262, (2003). Alternatively, and/or additionally, secondary image paths or phase detection pixels may be used to determine and adjust focus quality.

[0003] Unfortunately – and as is further known in the art – these strategies and techniques suffer from the metrics being ad hoc and not provably optimal, that they are not specific to scene content and may focus on regions with little content or interest, that calculation of the metric(s) is sometimes computationally intensive/expensive, and, in the case of phase detection systems, specialized hardware is required.

SUMMARY

[0004] An advance in the art is made according to aspects of the present disclosure directed to neural network focusing for imaging systems that advantageously overcomes these – and other – focusing difficulties associated with contemporary imaging systems. As will be shown and described, a neural network according to aspects of the present disclosure is trained based on know best focus scenes thereby optimizing focus/imaging performance and resulting quality.

[0005] Advantageously – and in sharp contrast to the prior art - the neural network employed in imaging systems according to aspects of the present disclosure may identify regions or targets of interest within a scene and selectively focus on such target(s). Additionally, such neural network may process subsampled or compressively sampled data directly thereby substantially reducing computational requirements as compared with prior-art contemporary systems, methods, and structures. Finally, such neural network may directly evaluate image quality on image data – thereby eliminating need for focus specific hardware.

BRIEF DESCRIPTION OF THE DRAWING

[0006] A more complete understanding of the present disclosure may be realized by reference to the accompanying drawing in which:

[0007] **FIG. 1** shows a schematic ray diagram of an illustrative optical arrangement according to aspects of the present disclosure;

[0008] **FIG. 2** shows a schematic diagram of an illustrative six layer neural network according to aspects of the present disclosure;

[0009] **FIG. 3** shows a pair of images captured at an optimal focus (Sharp Image - top) position and an off position (Blurred Image - bottom) according to aspects of the present disclosure;

[0010] **FIG. 4(A), FIG. 4(B), FIG. 4(C), FIG. 4(D), FIG. 4(E),** and **FIG. 4(F)** are a series of images illustrating model calibration using images captured with different focal position and displacement from focal position according to aspects of the present disclosure;

[0011] **FIG. 5** shows a schematic diagram of an illustrative seven layer neural network for predicting distance according to aspects of the present disclosure;

[0012] **FIG. 6** shows a schematic diagram of an illustrative discriminant six layer neural network according to aspects of the present disclosure;

[0013] **FIG. 7(A), FIG. 7(B),** and **FIG. 7(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 7(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 7(B)** and **FIG. 7(C)** – by comparison, the image of **FIG. 7(C)** is better than that of **FIG. 7(B)** and the discriminant predicts **FIG. 7(C)** is a focused image;

[0014] **FIG. 8(A), FIG. 8(B),** and **FIG. 8(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 8(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 8(B)** and **FIG. 8(C)** – by comparison, the image of **FIG. 8(C)** is better than that of **FIG. 8(B)** and the discriminant predicts **FIG. 8(C)** is a focused image;

[0015] **FIG. 9(A), FIG. 9(B),** and **FIG. 9(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 9(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 9(B)** and **FIG. 9(C)** – by comparison, the image of **FIG. 9(C)** is better than that of **FIG. 9(B)** and the discriminant predicts **FIG. 9(C)** is a focused image;

[0016] **FIG. 10(A), FIG. 10(B), FIG. 10(C), FIG. 10(D)**, and **FIG. 10(E)** is a series of images illustrating aspects of the present disclosure in which **FIG. 10(A)** is a

defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 10(B)** and **FIG. 10(C)** – by comparison, the image of **FIG. 10(C)** is better than that of **FIG. 10(B),** but the discriminant predicts **FIG. 10(C)** is still not a focused image, so a new movement is predicted based on **FIG. 10(C)** and two more images are captured – by comparison, image **FIG. 10(E)** is better than **FIG. 10(D)**, and the discriminant predicts image **FIG. 10(E)** is a focused image.

[0017] **FIG. 11** shows a schematic diagram of an illustrative seven layer neural network for compressively sensing according to aspects of the present disclosure;

[0018] **FIG. 12(A), FIG. 12(B), FIG. 12(C)** and **FIG. 12(D)** is a series of images illustrating aspects of the present disclosure in which **FIG. 12(A)** is a defocused image which is compressively sensed to a four channel tensor by a random matrix wherein each channel is shown in **FIG. 12(B),** the network leans from these images, any movement required and image **FIG. 12(C)** and **FIG 12(D)** is classified focused by the discriminator;

[0019] **FIG. 13** shows a schematic diagram of an illustrative neural network for salience detection according to aspects of the present disclosure;

[0020] **FIG. 14(A), FIG. 14(B),** and **FIG. 14(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 14(A)** is an original image, **FIG. 14(B)** shows a generated defocused image and **FIG. 14(C)** shows the label;

[0021] **FIG. 15(A), FIG. 15(B),** and **FIG. 15(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 15(A)** shows the first defocused image captured by a camera, **FIG. 15(B)** shows the predicted saliency map, and **FIG. 15(C)** shows the final output image with bounding box showing the block that is selected as the target;

[0022] **FIG. 16(A), FIG. 16(B),** and **FIG. 16(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 16(A)** shows the first defocused image captured by a camera, **FIG. 16(B)** shows the predicted saliency map, and **FIG. 16(C)** shows the final output image with bounding box showing the block that is selected as the target; and

[0023] **FIG. 17(A), FIG. 17(B),** and **FIG. 17(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 17(A)** shows the first defocused image captured by a camera, **FIG. 17(B)** shows the predicted saliency map, and **FIG. 17(C)** shows the final output image with bounding box showing the block that is selected as the target.

[0024] The illustrative embodiments are described more fully by the Figures and detailed description. Embodiments according to this disclosure may, however, be embodied in various forms and are not limited to specific or illustrative embodiments described in the drawing and detailed description.

DESCRIPTION

[0025] The following merely illustrates the principles of the disclosure. It will thus be appreciated that those skilled in the art will be able to devise various arrangements which, although not explicitly described or shown herein, embody the principles of the disclosure and are included within its spirit and scope.

[0026] Furthermore, all examples and conditional language recited herein are intended to be only for pedagogical purposes to aid the reader in understanding the principles of the disclosure and the concepts contributed by the inventor(s) to furthering the art and are to be construed as being without limitation to such specifically recited examples and conditions.

[0027] Moreover, all statements herein reciting principles, aspects, and embodiments of the disclosure, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended

that such equivalents include both currently known equivalents as well as equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure.

[0028] Thus, for example, it will be appreciated by those skilled in the art that any block diagrams herein represent conceptual views of illustrative circuitry embodying the principles of the disclosure.

[0029] Unless otherwise explicitly specified herein, the FIGs comprising the drawing are not drawn to scale.

[0030] Finally, it is noted that the use herein 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.

[0031] By way of some additional background, we note that a primary task of camera focus control is to set a sensor plane at which light from an object converges. **FIG. 1** shows a schematic ray diagram of an illustrative optical arrangement according to aspects of the present disclosure.

[0032] As shown in **FIG. 1**, the light from the object converges at Z_0 on optical axis, while the sensor plane can move between Z_{min} and Z_{max}. As such focus control requires finding Z_0, and moving the sensor plane accordingly. To achieve this, existing

auto-focus algorithms set the sensor plane at different positions between Z_{min} and Z_{max} and determine its optimal position by comparing the captured image quality with respect to one or more of the metrics previously mentioned.

[0033] According to aspects of the present disclosure – and in further contrast to the prior art - instead of capturing a large number of images at different positions to search for an optimal one, only two gray-scale images are captured or converted from RGB images, denoted as I_{Z_1} and I_{Z_2}, wherein the subscripts represent the position of the sensor plane. Note that as used herein, 'image' describes a gray-scale image obtained from a camera or other imaging apparatus/structure.

[0034] While Z_1 and Z_2 can be any number between Z_{min} and Z_{max}, the displacement between Z_1 and Z_2 is fixed and defined by the following relationship:

$$Z_1 - Z_2 = Z_d.$$

[0035] Then two image blocks, B_{Z_1} and B_{Z_2}, containing the object of interest are cropped from the two images, I_{Z_1} and I_{Z_2}, at the same position. The image blocks are then normalized to be between 0 and 1:

$$P = \frac{B - B_{min}}{B_{max} - B_{min}}$$

[0036] The neural-network-based focus position prediction algorithm according to aspects of the present disclosure - denoted as f_{Z_d} - learns from the two normalized blocks the displacement between Z_1 and Z_0, according to:

$$f_{Z_d}(P_{Z_1}, P_{Z_2}) = Z_0 - Z_1$$

[0037] Consequently – and as will now be readily understood by those skilled in the art - the optimal focus position, Z_0, can be achieved by moving the sensor plane accordingly.

Neural Network Structure

[0038] **FIG. 2** shows a schematic diagram of an illustrative six layer neural network according to aspects of the present disclosure. With reference to that figure we note that such neural network according to the present disclosure receives as input two normalized image blocks, P_{Z_1} and P_{Z_2}, and predicts the displacement between Z_1 and Z_0. As viewed from left to right, the first two arrows represent convolutional layers and subsequent four arrows represent fully-connected layers. Two image blocks of size 256×256 are first stacked, forming a $256 \times 256 \times 2$ input tensor.

[0039] As may be further observed from the figure, the illustrative neural network includes six consecutive layers. The input tensor is first applied (fed) into a convolutional layer. Illustrative filter size, number of filters and stride for this layer are 8, 64 and 8 respectively, and the activation function for this layer is rectified linear unit (ReLu). The dimension of the output of this layer is $32 \times 32 \times 64$.

[0040] The second illustrative layer is another convolutional layer having illustrative filter size, number of filters and stride being 4, 16 and 4 respectively. The activation function for this layer is ReLu, and the dimension of the output is $8 \times 8 \times 16$. The four subsequent illustrative layers are fully-connected layers, and the dimensions of the four fully-connected layers are 1024, 512, 10 and 1 respectively. ReLu activation is applied to the first three fully-connected layers.

Defocus Model

[0041] At this point we note that focus control is an imaging-system-specific problem, so each imaging system (i.e., camera) requires a unique network trained on data for that specific imaging system. Advantageously, systems, methods, and structures according to aspects of the present disclosure may include one or more mechanism(s) to construct a defocus model for the specific imaging system, and the model can be used to generate training data from existing image dataset.

[0042] **FIG. 3** shows a pair of images captured at an optimal focus (Sharp Image - top) position and an off position (Blurred Image - bottom) according to aspects of the present disclosure. The defocus process can be modeled as image blur followed by image scaling defined by:

$$Gamma^{-1}(I_z) = imscale(Gamma^{-1}(I_{Z_0}) * h, \alpha),$$

where h is the image blur filter, α is the image scaling factor and $Gamma^{-1}(x)$ is the inverse process of gamma correction defined by:

$$Gamma^{-1}(x) = x^{\gamma}.$$

[0043] This inverse process compensates for non-linear operation, *i.e.* gamma correction, in the image signal processor (ISP), where γ is predefined for a specific camera. The present disclosure assumes h is a circular averaging filter, which is uniquely determined by the averaging radius, r. Both r and α are determined by the optimal focus position, Z_0, and the displacement from the optimal focus position, $Z - Z_0$.

$$r, \alpha = g(Z_0, Z - Z_0)$$

[0044] To calibrate the model, systems, methods, and structures according to aspects of the present disclosure require images captured at different focal position and displacement from focal position, *i.e.* Z_0 and $Z - Z_0$. An object is installed with different distance, d_0, in front of the camera (see **FIG. 1**), and each d_0 corresponds to a unique Z_0 and I_{Z_0}.

[0045] For each Z_0, by moving the sensor plane between Z_{min} and Z_{max}, images with different Z can be captured, and each pair of I_z and I_{Z_0} are used to calibrate $g(Z_0, Z - Z_0)$. – with reference to **FIG.4(A), FIG.4(B), FIG.4(C), FIG.4(D), FIG.4(E),** and **FIG.4(F), -** the overall illustrative process may be described as follows

1. First, an image block from $Gamma^{-1}(I_{z0})$ is cropped (see the bounding box in **FIG. 4 (A)**), denoted as J_{Z_0} (see **FIG. 4 (B)**). J_{Z_0} should be included in an object that all the pixels have the same distance d_0.

2. For each r, a blurred image is generated by convolving J_{z0} and $h(r)$. (see **FIG.4(C)**).

3. A block $K_{Z_0,h(r)}$ is cropped from the blurred image (see the bounding box in **FIG.4(C)** and **FIG. 4(D)**). This step is to remove pixels that should have been blurred by the pixels outside the J_{Z_0}.

4. For each α, a scaled image is generated from $K_{Z_0,h(r)}$. (**FIG. 4(E)**)
5. A score for each pair of r and α is calculated by computing the maximum normalized cross-correlation between the scaled $K_{Z_0,h(r)}$ and all blocks that have the same dimension in $Gamma^{-1}(I_z)$. (see, **FIG.4(F)**)

6. The r and α corresponding to the maximum score are the calibrated model parameters.

[0046] After calibration, a defocused image can be generated by:

$$d_z = Gamma(imscale(Gamma^{-1}(d_{Z_0}) * h(r(Z_0, Z - Z_0)), \alpha(Z_0, Z - Z_0)),$$

where d_{Z_0} is a clear image from existing image dataset which is assumed to be the image at Z_0, d_z is the simulated defocused image with respect to Z_0 and Z, and $Gamma(x)$ is defined by:

$$Gamma(x) = x^{\frac{1}{\gamma}}.$$

Training Data

[0047] Once the model is calibrated, training data can be generated from existing image dataset(s). First, the number of training data, N, and Z_d are chosen. For each grayscale image d_i:

1. An optimal focus position, $Z_{i,0}$, and a sensor plane position, $Z_{i,1}$, are randomly generated:
$$Z_{i,0} \in [Z_{min}, Z_{max}], Z_{i,1} \in [Z_{min}, Z_{max} - Z_d]$$

2. Two defocused images are generated:

$$d_{i,Z_1} = Gamma\{imscale[Gamma^{-1}(d_i) * h(r(Z_{i,0}, Z_{i,1} - Z_{i,0})), \alpha(Z_{i,0}, Z_{i,1} - Z_{i,0})]\},$$
$$d_{i,Z_2} = Gamma\{imscale[Gamma^{-1}(d_i) * h(r(Z_{i,0}, Z_{i,2} - Z_{i,0})), \alpha(Z_{i,0}, Z_{i,2} - Z_{i,0})]\},$$

where $Z_{i,2} = Z_{i,1} + Z_d$.

3. Two image blocks, B_{i,Z_1} and B_{i,Z_2}, containing the object of interest are cropped from the two images, d_{i,Z_1} and d_{i,Z_2}, at the same position. The image blocks are then normalized to be between 0 and 1 according to the following:

$$\bar{d} = \frac{B - B_{min}}{B_{max} - B_{min}}.$$

[0048] Each data comprises two input images, $\bar{d}_{i,Z_1}$ and $\bar{d}_{i,Z_2}$, and the corresponding label $Z_{i,0} - Z_{i,1}$.

[0049] The neural network is trained on the simulated dataset by minimizing the mean square error, *i.e.*

$$L = \frac{1}{N}\sum_{i=1}^{N}[f(\bar{d}_{i,Z_1}, \bar{d}_{i,Z_2}) - Z_{i,0} + Z_{i,1}]^2,$$

and

$$f_{Z_d} = arg\min_f \frac{1}{N}\sum_{i=1}^{N}[f(\bar{d}_{i,Z_1}, \bar{d}_{i,Z_2}) - Z_{i,0} + Z_{i,1}]^2$$

Bayer Data

[0050] With a digital camera, data directly captured from the scene is the raw data in Bayer format. For cameras with access to its raw data, systems, methods, and structures according to aspects of the present disclosure may advantageously be modified to perform on raw data. To calibrate the model, each raw data, a one channel intensity map, is first interpolated to generate an RGB image. Then the same calibration process is applied. Note that we assume that the defocus model is the same for three channels.

[0051] To generated simulated training data, each raw data is first interpolated. Then the defocus model is applied to each channel, and the defocused raw data is generated

by mosaicking the defocused RGB image. The same network structure and training strategy can be applied to train the model.

Multi-Step Focus

[0052] Advantageously according to aspects of the present disclosure - and instead of analyzing multiple images to find the optimal focus position - one defocused image is sufficient to determine the absolute distance between the optimal focus position, $abs(Z - Z_0)$, and the current sensor plane, Z.

[0053] According to aspects of the present disclosure, a multi-step focus mechanism may be employed. Operationally, when a defocused image, I_z, is captured, a neural network, denoted as $f_{distance}$, is applied to find the distance between the optimal focus position and the current position, $Z_d = abs(Z - Z_0)$. Then, two images are captured by moving the sensor plane in different directions, *i.e.* I_{Z-Z_d} by moving the sensor plane to $Z - Z_d$ and I_{Z+Z_d} by moving the sensor plane to $Z + Z_d$, and the two images are compared to find the right moving direction. By introducing a discriminant, denoted as $f_{discriminant}$, to judge if the image is focused, this solution can adjust the focus until a clear image is captured.

Network Details

$f_{distance}$

[0054] **FIG. 5** shows a schematic diagram of an illustrative seven layer neural network for predicting distance Z_d according to aspects of the present disclosure.

[0055] The input to the network is a 512×512 block from the defocused image. The network comprises seven consecutive layers. The input image block is first fed into a convolutional layer. The illustrative filter size, number of filters and stride for this layer are 8, 64 and 8 respectively, and the activation function for this layer is rectified linear unit (ReLu). The illustrative dimension of the output of this layer is $64 \times 64 \times 4$.

[0056] The second layer and the third layer are illustratively convolutional layers with filter size/number of filters/stride being 4/8/4 and 4/8/4 respectively. The activation function is ReLu, and the dimensions of the output are $16 \times 16 \times 8$ and $4 \times 4 \times 8$.

[0057] The four subsequent layers are illustratively fully-connected layers, and the dimensions of the four fully-connected layers are 1024, 512, 10 and 1 respectively. Leaky ReLu activation is applied to the first three fully-connected layers.

$f_{discriminant}$

[0058] **FIG. 6** shows a schematic diagram of an illustrative discriminant six layer neural network according to aspects of the present disclosure. The input to the network is a 512×512 block from image.

[0059] As may be observed, there are five consecutive layers in this illustrative network. The first two layers are illustratively convolutional layers, with filter size/number of filters/stride being 8/1/8 and 8/1/8 respectively. The activation function for these two layers is ReLu, and the dimensions of the illustrative output of these two layers are $64 \times 64 \times 1$ and $8 \times 8 \times 1$.

[0060] The third illustrative layer is a fully-connected layer, followed by the ReLU activation, and the dimension of this layer is 10. The fourth layer is a dropout layer with rate 0.5. Last layer is illustratively a fully-connected layer followed by softmax activation which indicates the category of the input block.

Algorithm Details:

[0061] With the foregoing in place, we may now describe an entire illustrative process according to aspects of the present disclosure as follows:

1. An image is captured at Z, denoted I_z, and a 512×512 block is cropped and normalized, denoted J_z.

2. Determine if the image is focused by $f_{discriminant}$. If YES, stop the algorithm and output Z; If NOT, continue the algorithm.

3. Predict the distance from the optimal focus position
$$Z_d = \boldsymbol{f}_{\boldsymbol{distance}}(J_z)$$

4. Compute the two possible positions $Z + Z_d$ and $Z - Z_d$, and determine which is/are feasible (between the Z_{min} and Z_{max}).

5. If only one position is feasible, update Z and return to step 1.

6. If both positions are feasible, capture the images at $Z + Z_d$ and $Z - Z_d$, denoted I_{Z+Z_d} and I_{Z-Z_d}. Predict the distance between the optimal focus position and the focus position of the two images
$$Z_{d1} = \boldsymbol{f}_{\boldsymbol{distance}}(J_{Z+Z_d}),\ Z_{d2} = \boldsymbol{f}_{\boldsymbol{distance}}(J_{Z-Z_d})$$

7. If $Z_{d1} < Z_{d2}$:

    1. Update $Z = Z + Z_d$ and $Z_d = Z_{d1}$

    2. Determine if the image I_{Z+Z_d} is focused by $f_{discriminant}$. If YES, stop the algorithm and output Z; If NOT, return to step 4.

    If $Z_{d1} > Z_{d2}$:

    1. Update $Z = Z - Z_d$ and $Z_d = Z_{d2}$

    2. Determine if the image I_{Z-Z_d} is focused by $f_{discriminant}$. If YES, stop the algorithm and output Z; If NOT, return to step 4.

8. If none of the positions is feasible, update Z with a small step change
$$Z = Z + \hat{Z},$$
and return to step 1.

Examples

[0062] **FIG. 7(A), FIG. 7(B),** and **FIG. 7(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 7(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 7(B)** and **FIG. 7(C)** – by comparison, the image of **FIG. 7(C)** is better than that of **FIG. 7(B)** and the discriminant predicts **FIG. 7(C)** is a focused image.

[0063] **FIG. 8(A), FIG. 8(B),** and **FIG. 8(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 8(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 8(B)** and **FIG. 8(C)** – by comparison, the image of **FIG. 8(C)** is better than that of **FIG. 8(B)** and the discriminant predicts **FIG. 8(C)** is a focused image.

[0064] **FIG. 9(A), FIG. 9(B),** and **FIG. 9(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 9(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 9(B)** and **FIG. 9(C)** – by comparison, the image of **FIG. 9(C)** is better than that of **FIG. 9(B)** and the discriminant predicts **FIG. 9(C)** is a focused image.

[0065] **FIG. 10(A), FIG. 10(B), FIG. 10(C), FIG. 10(D)**, and **FIG. 10(E)** is a series of images illustrating aspects of the present disclosure in which **FIG. 10(A)** is a defocused image captured at the beginning of the process, then the network predicts movement and generates two possible positions illustrated in **FIG. 10(B)** and **FIG. 10(C)** – by comparison, the image of **FIG. 10(C)** is better than that of **FIG. 10(B),** but the discriminant predicts **FIG. 10(C)** is still not a focused image, so a new movement is predicted based on **FIG. 10(C)** and two more images are captured – by comparison, image **FIG. 10(E)** is better than **FIG. 10(D)**, and the discriminant predicts image **FIG. 10(E)** is a focused image.

Compressively Sensed Image Data

[0066] Compressive sensing is a potential technique in image systems because it achieves sensor level compression. While traditional methods require analyzing images to determine the focus position, the present disclosure determines the focus from compressed data. While the first convolutional layer in the network provides a sensing matrix for compressive sensing, and this sensing matrix is trained to yield the best focus results, a more general sensing matrix is tested which proves that the present solution can be applied to compressed data.

[0067] **FIG. 11** shows a schematic diagram of an illustrative seven layer neural network for compressively sensing according to aspects of the present disclosure. As may be observed from that figure, the basic network structure is substantially the same as that shown previously, wherein the first illustrative convolutional layer is randomly initialized and is not trainable. This is equivalent to generating a random sensing matrix for compressive sensing, and here the measurement rate is 0.0625.

[0068] **FIG. 12(A), FIG. 12(B), FIG. 12(C)** and **FIG. 12(D)** is a series of images illustrating aspects of the present disclosure in which **FIG. 12(A)** is a defocused image which is compressively sensed to a four channel tensor by a random matrix wherein each channel is shown in **FIG. 12(B),** the network leans from these images, any movement required and image **FIG. 12(C)** and **FIG 12(D)** is classified focused by the discriminator. Note that image of **FIG. 12(B)** is not downsampled from the image of **FIG. 12(A).**

Saliency-based Focus

[0069] In imaging systems, a focus module is controlled to "best image" (i.e., determine a best image) a target object/target object plane. However, since the real world is 3-dimentional, an indication of the target is required in focus control. In current imaging systems, such as digital cameras or cellphone cameras, the location of the target is: 1) by default the central of the image, or 2) manually indicated by users. Some smart systems may adjust the focus module to best capture faces by first detecting the faces in the image, but the detection requires a clear image, which is based on a general focus method per se. Visual saliency detection has been explored for decades, and advantageously according to aspects of the present disclosure - it may be combined with camera focus control algorithm, providing a saliency-based focus control mechanism.

[0070] As we shall show and describe, the main workflow of saliency-based focus according to the present disclosure is like the workflow disclosed herein for Multi-step Focus, wherein the location of the block is determined by a saliency detection network from the first defocused image I_z.

[0071] **FIG. 13** shows a schematic diagram of an illustrative neural network for salience detection according to aspects of the present disclosure. Operationally, the image is resized to 300×400 before being fed into the network. The input image is first fed into three separate convolutional layers, the filter sizes for the three convolutional layers are 40, 10 and 3 individually, and the number of filters/stride for each layer are 16/1.

[0072] The output of the three layers are then concatenated, forming a $300 \times 400 \times 48$ tensor, denoted in the figure as CONV1. Following are two convolutional layers, with filter size/strides being 10/10 for both layers. The number of filters for the convolutional layers are 16 and 32 individually, and the output dimension of two layers are $30 \times 40 \times 16$ and $3 \times 4 \times 32$, denoted in the figure as CONV2 and CONV3.

[0073] Next, CONV3 is reshaped to a 384×1 vector, followed by a fully-connected layer with dimension 384×1, which is again reshaped to a $3 \times 4 \times 32$ tensor, denoted CONV4. CONV3 and CONV4 are then concatenated to form a $3 \times 4 \times 64$ tensor, followed by a convolutional layer with filter size/number of filters/stride being 3/64/1.

[0074] A deconvolutional layer is then applied with filter size/number of filters/stride being 20/16/10, and the dimension of the output tensor, denoted CONV5, is $30 \times 40 \times 16$. CONV2 and CONV5 are then concatenated to form a $30 \times 40 \times 32$ tensor, followed by a convolutional layer with filter size/number of filters/stride being 5/64/1.

[0075] Then another deconvolutional layer is applied with filter size/number of filters/stride being 20/32/10, and the dimension of the output tensor, denoted CONV6, is $300 \times 400 \times 32$.

[0076] Finally, CONV1, CONV6 and the input image are concatenated to form a $300 \times 400 \times 83$ tensor, followed by four consecutive convolutional layers. The filter size/number of filters/stride for the three layers are 11/64/1, 7/32/1, 3/16/1 and 1/1/1, and the output of the last layer is the predicted saliency map. ReLU activation is applied to all

convolutional, deconvolutional and fully-connected layers except for the last layer, which is activated by Sigmoid function.

Data Generation

[0077] There exist several labeled datasets (i.e., MSRA10K*, MSRA-B**, PASCALS***) for saliency detection, but no publicly available dataset provides labeled defocused image for saliency detection. To prepare training data, the image defocus model calibrated on a specific camera are used, and process according to aspects of the present disclosure is described as:

1. Obtain an image and its label from existing datasets.
2. Resize the image to the size of the camera image.
3. Regard the resized image as a focused image and generate a defocused image using the calibrated model.
4. Resize the defocused image and its corresponding label to 300×400. An example (image from MSRA-B**) is shown in **FIG. 14(A)**, **FIG. 14(B)**, and **FIG. 14(C)**. Image **FIG. 14(A)** shows the original image, image **FIG. 14(B)** shows the generated defocused image, and image **FIG. 14(C)** shows the label. Each training data consists of a defocused image I and a label image L.

[0078] The network $f_{Saliency}$ is trained on the training dataset by minimizing the mean square error, *i.e.*

$$f_{Saliency} = arg\min_{f} \frac{1}{N} \sum_{i=1}^{N} [f(I_i) - L_i]^2.$$

Block Selection:

[0079] When the saliency map is predicted by the network, the next step is to select the block to perform the focus prediction network on. To achieve that, the saliency map is first resized to the size of the original image, *i.e.* image captured by the camera. Then the block is determined by finding a 512×512 block that achieves the highest average intensity:

$$B^* = arg\max_{B} \sum_{I_{i,j} \in B} S_{i,j},$$

where B represents a 512×512 block from the image, $I_{i,j}$ is a pixel in the image, and $S_{i,j}$ represents the saliency map. Once the block is determined, the focus is predicted from this block.

Examples

[0080] **FIG. 15(A), FIG. 15(B),** and **FIG. 15(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 15(A)** shows the first defocused image captured by a camera, **FIG. 15(B)** shows the predicted saliency map, and **FIG. 15(C)** shows the final output image with bounding box showing the block that is selected as the target.

[0081] **FIG. 16(A), FIG. 16(B),** and **FIG. 16(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 16(A)** shows the first defocused image captured by a camera, **FIG. 16(B)** shows the predicted saliency map, and **FIG. 16(C)** shows the final output image with bounding box showing the block that is selected as the target.

[0082] **FIG. 17(A), FIG. 17(B),** and **FIG. 17(C)** is a series of images illustrating aspects of the present disclosure in which **FIG. 17(A)** shows the first defocused image captured by a camera, **FIG. 17(B)** shows the predicted saliency map, and **FIG. 17(C)** shows the final output image with bounding box showing the block that is selected as the target.

[0083] At this point, while we have presented this disclosure using some specific examples, those skilled in the art will recognize that our teachings are not so limited. Accordingly, this disclosure should be only limited by the scope of the claims attached hereto.

Claims:

1. A neural-network-based focusing method for an imaging system comprising:
   a) providing two gray-scale images, each individual one of the gray-scale images captured at a different position of a sensor plane of the imaging system;
   b) cropping two image blocks, one from each of the two gray-scale images provided and cropped from the same relative position to their respective gray-scale images, each individual one of the cropped image blocks including an object of interest; and
   c) predicting a focus position through the effect of the neural network operating on the two image blocks.

2. The neural-network-based focusing method according to claim 1 further comprising:
   d) moving the sensor plane and repeating steps a) – c).

3. The neural-network-based focusing method according to claim 2 further comprising:
   normalizing each individual one of the two cropped image blocks such that they are between 0 and 1.

4. The neural-network-based focusing method according to claim 3 further comprising:
   determining, through the effect of the neural-network, a displacement between the two provided gray-scale images from the two normalized blocks.

5. The neural-network-based focusing method according to claim 4 further comprising:
   training the neural network on training data for the imaging system, specifically.

6. The neural-network-based focusing method according to claim 5 further comprising:
   constructing a defocus model for the imaging system, specifically; and
   generating the imaging system specific training data. training the neural network.

7. The neural-network-based focusing method according to claim 6 further wherein the defocus model is generated from a defocus process, said defocus process modeled as image blur followed by image scaling.

8. The neural-network-based focusing method according to claim 7 wherein the defocus process is defined by

$$Gamma^{-1}(I_z) = imscale(Gamma^{-1}(I_{Z_0}) * h, \alpha),$$

where h is the image blur filter, α is the image scaling factor and $Gamma^{-1}(x)$ is the inverse process of gamma correction defined by:

$$Gamma^{-1}(x) = x^{\gamma}.$$

9. The neural-network-based focusing method according to claim 1 wherein the gray-scale images are derived from raw imaging system data in Bayer format, wherein each raw data comprises a one channel intensity map and interpolated to generate an RGB image.

10. The neural-network-based focusing method according to claim 1 wherein the gray-scale images are derived from moving the sensor plane in different directions and then determining a correct moving direction.

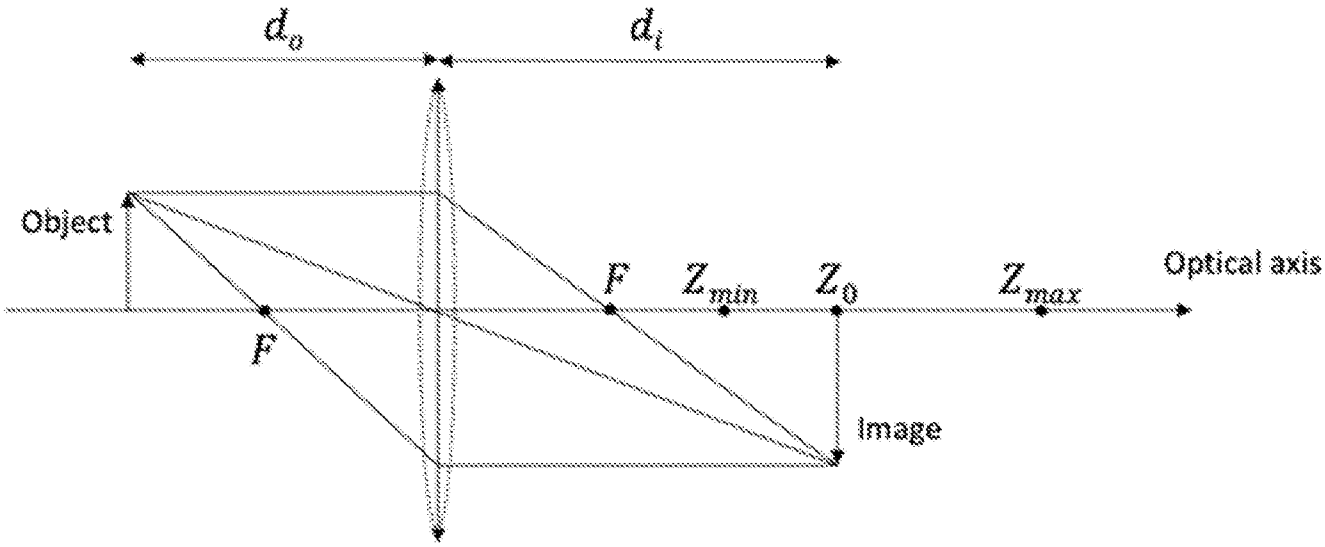


FIG. 1

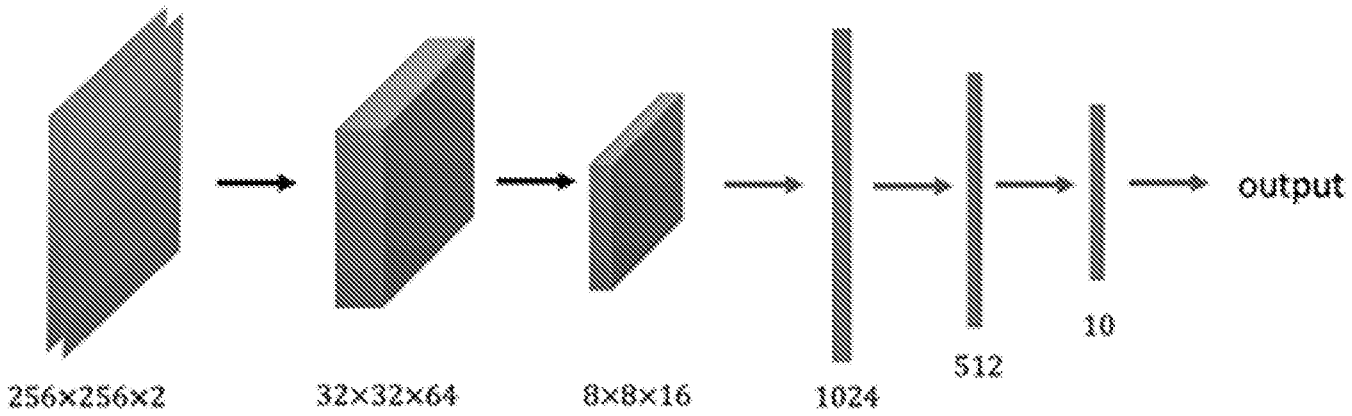
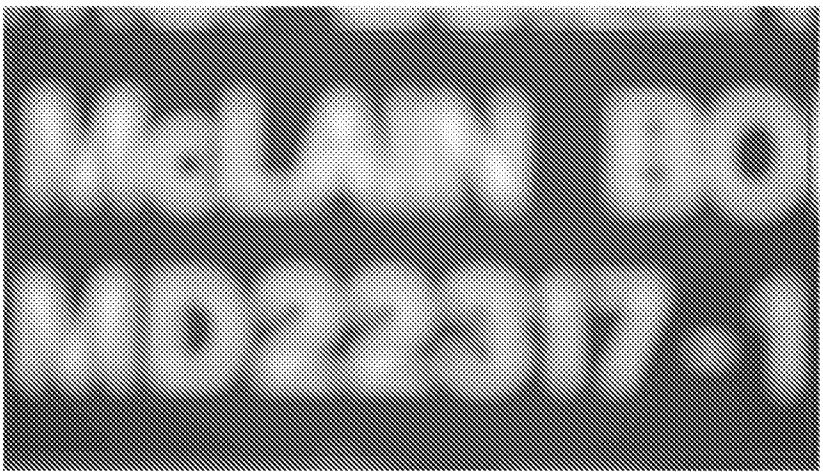


FIG. 2

Sharp Image



Blurred Image

FIG. 3

$Gamma^{-1}(I_{Z_0})$

FIG. 4(A)

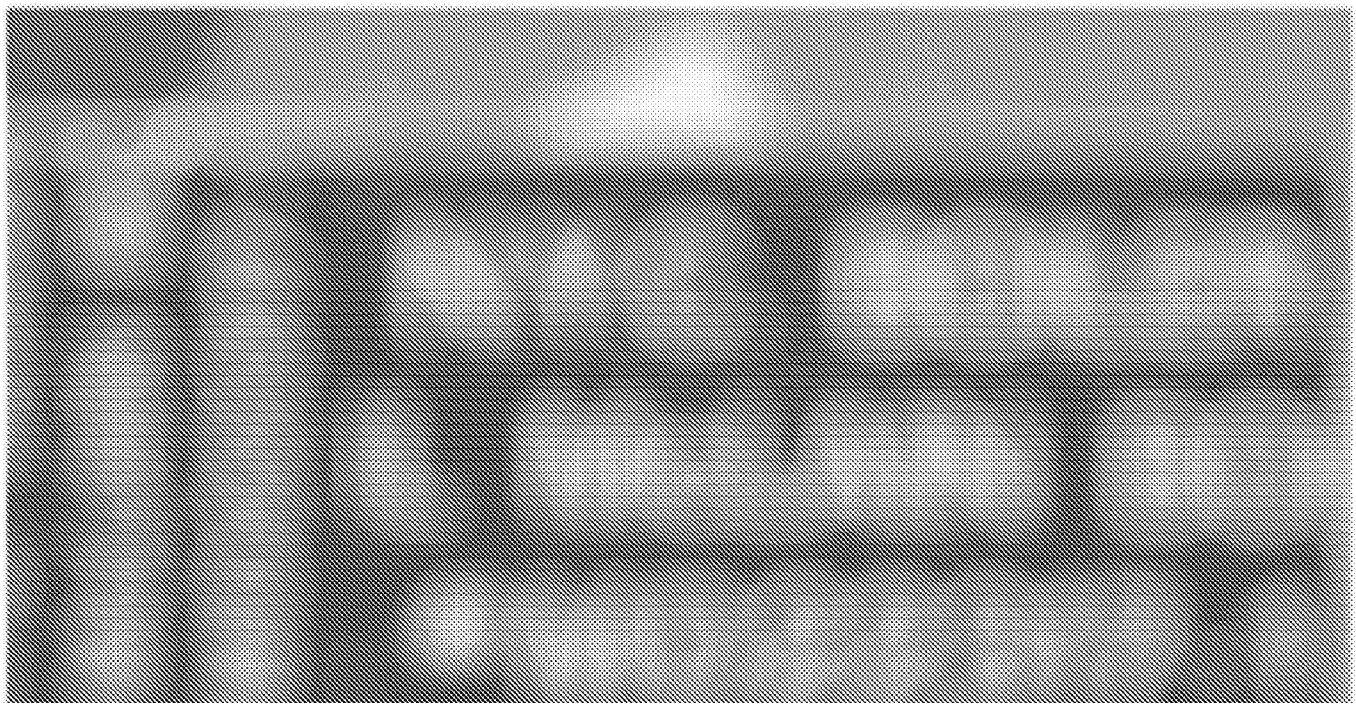
J_{z_0}

FIG. 4(B)

$J_{z_0} * h(r)$

Blurred Image

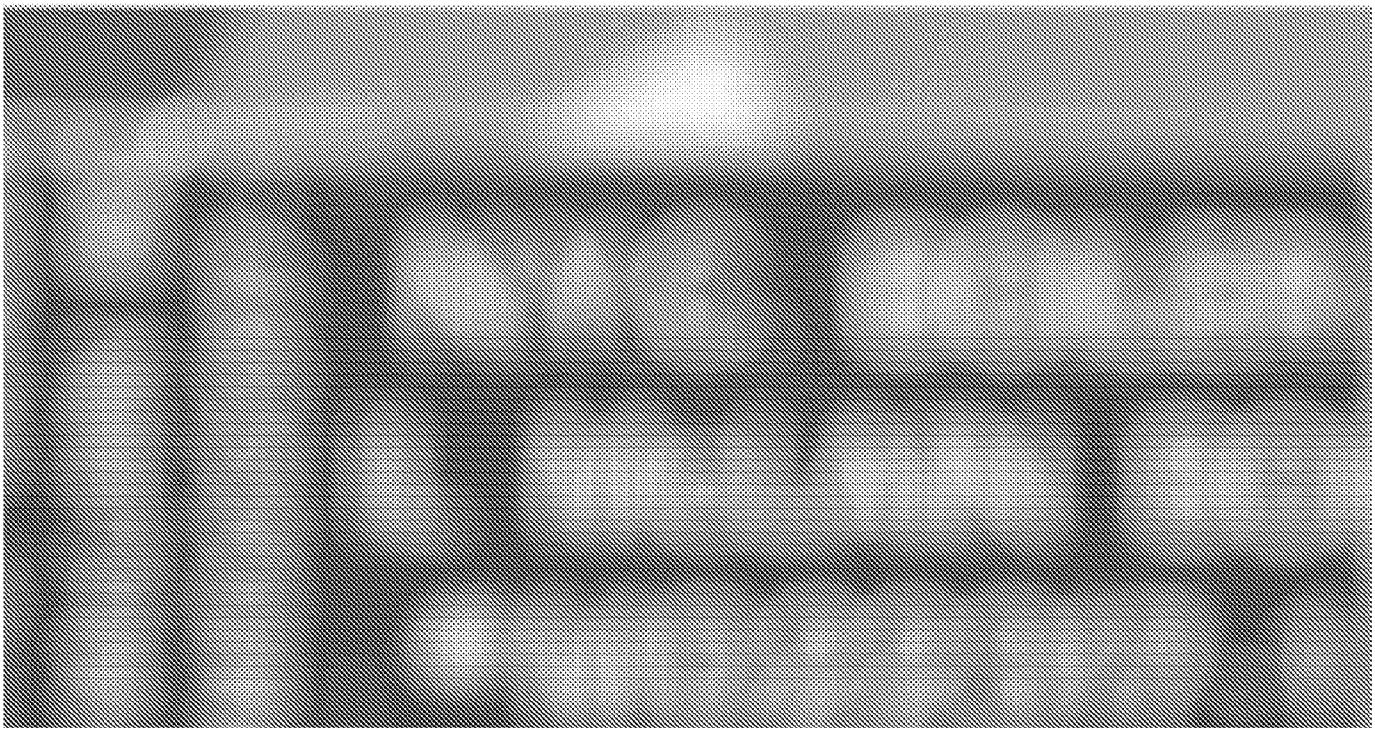
FIG. 4(C)



$K_{Z_0,h(r)}$

Blurred Image

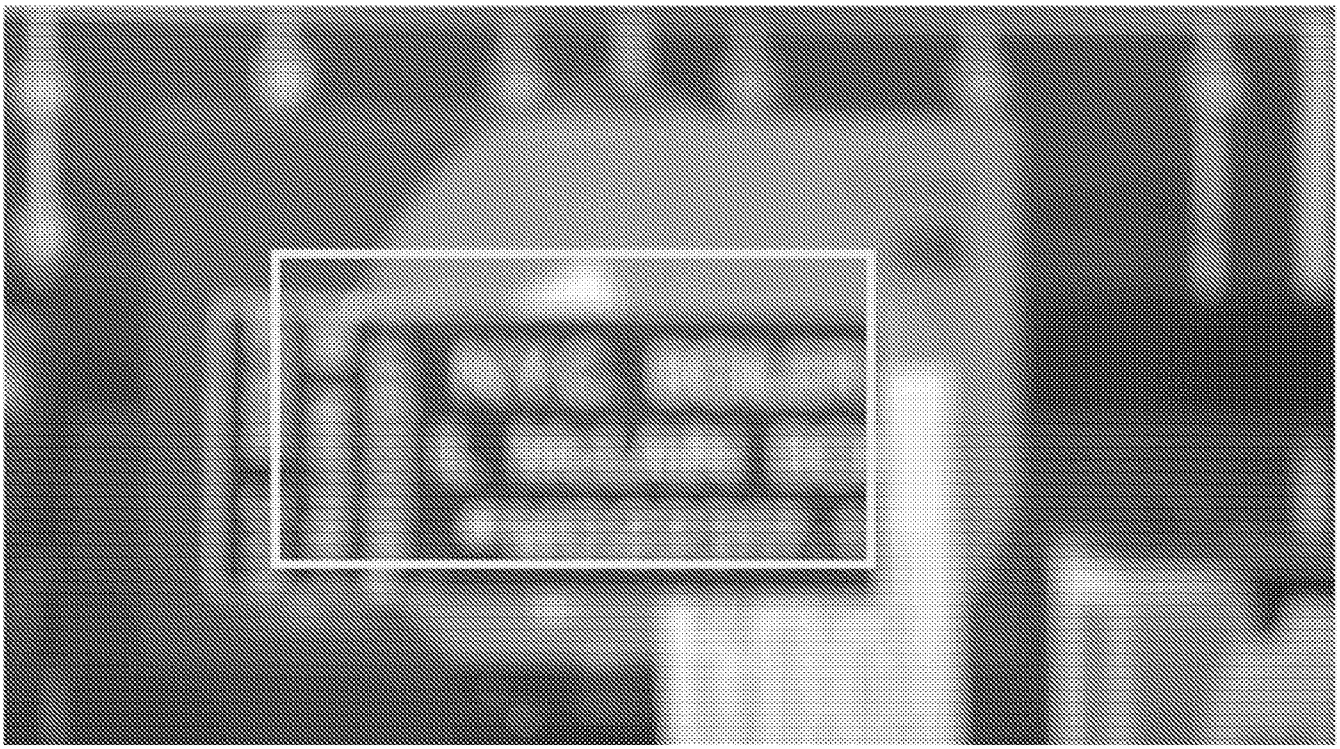
FIG. 4(D)



$imscale(K_{Z_0,h(r)}, \alpha)$

Blurred Image

FIG. 4(E)



$Gamma^{-1}(I_Z)$

Blurred Image

FIG. 4(F)

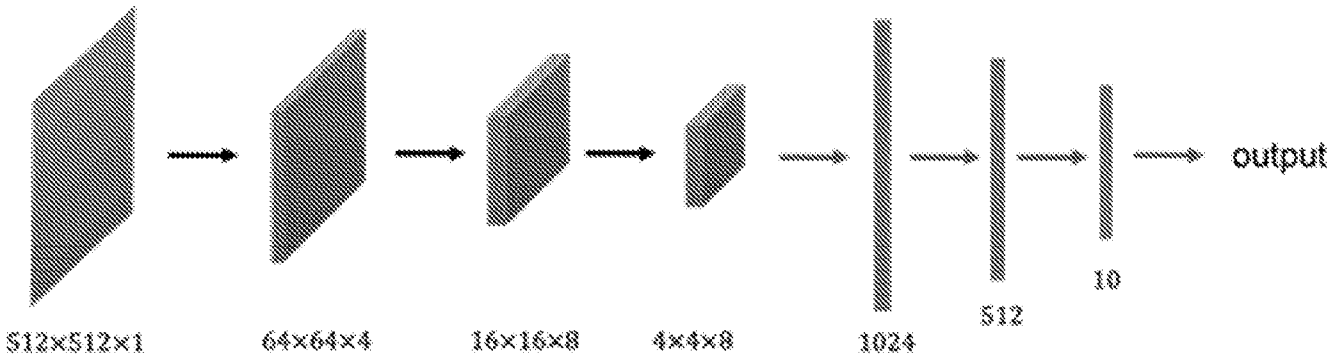


FIG. 5

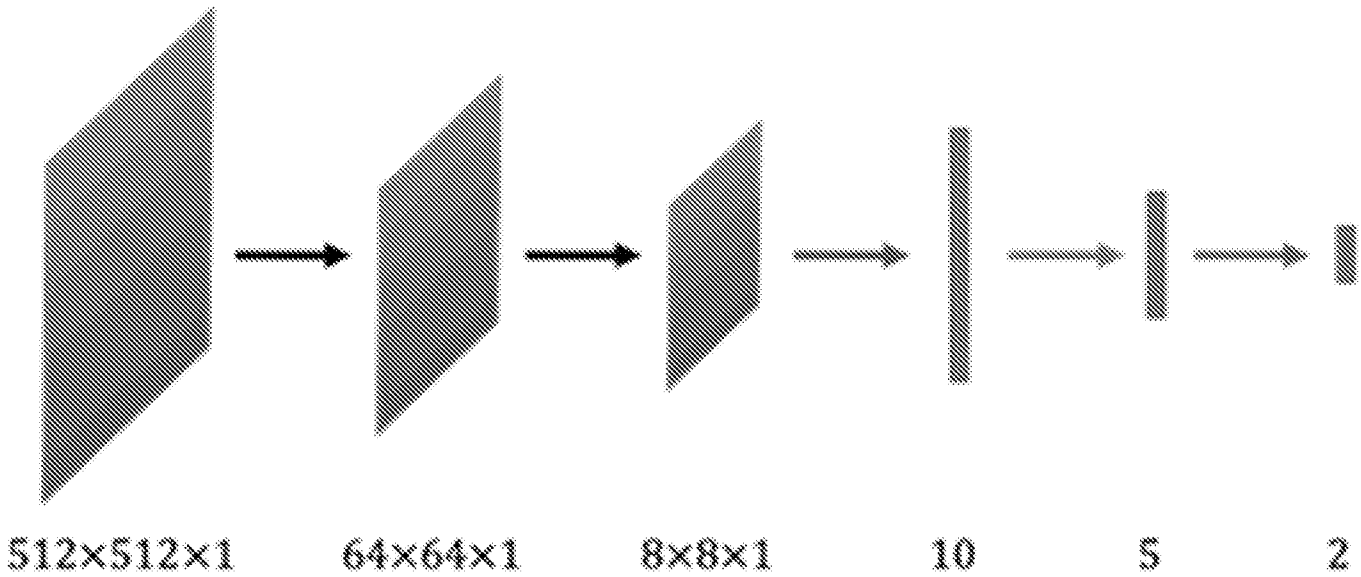
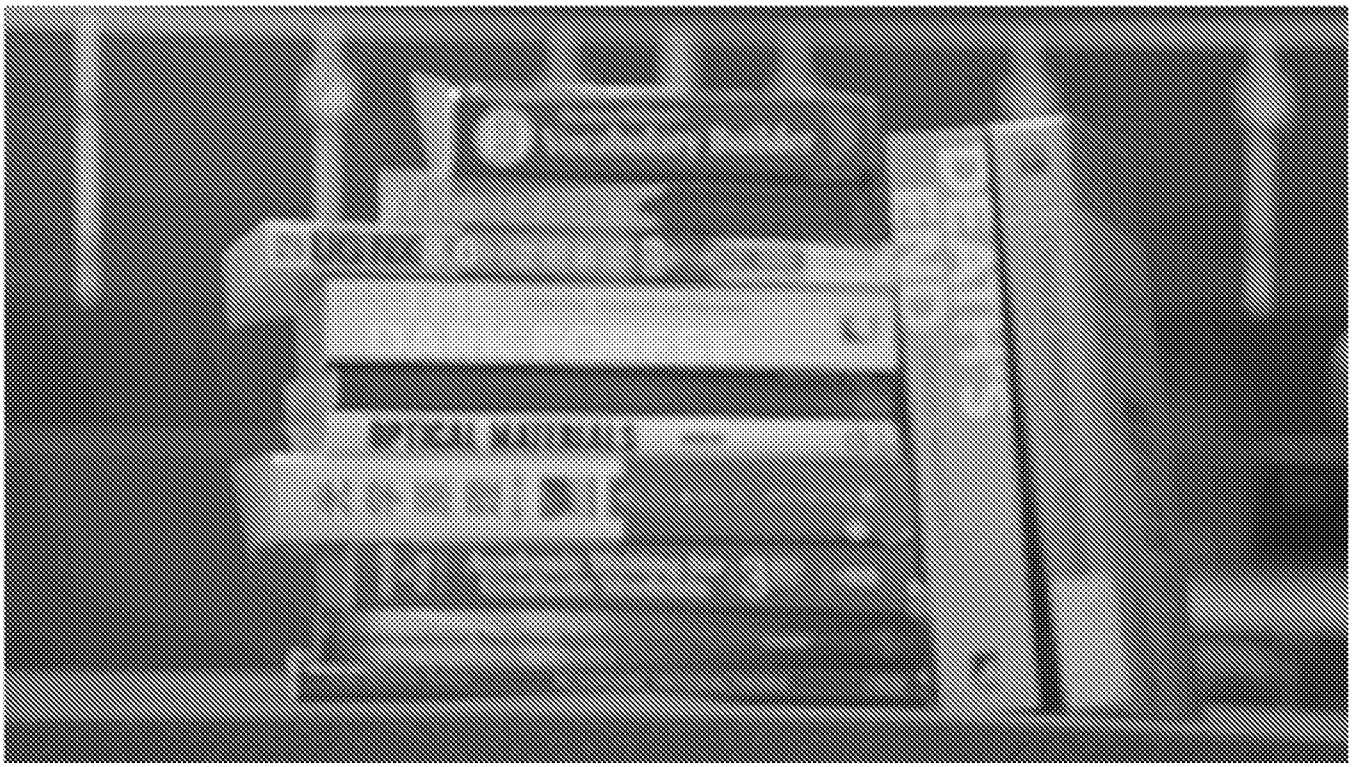


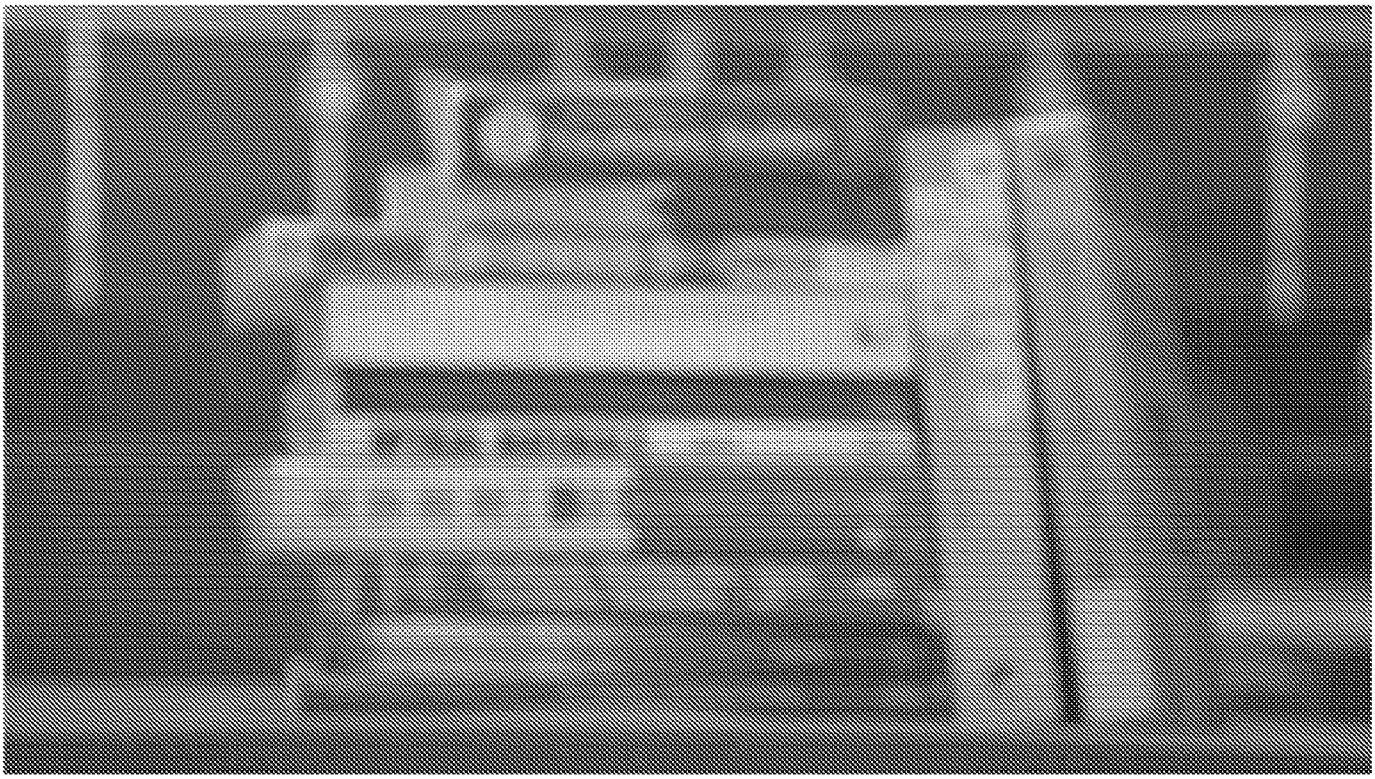
FIG.6



Example 1 (object distance 2.5m)

Out of Focus

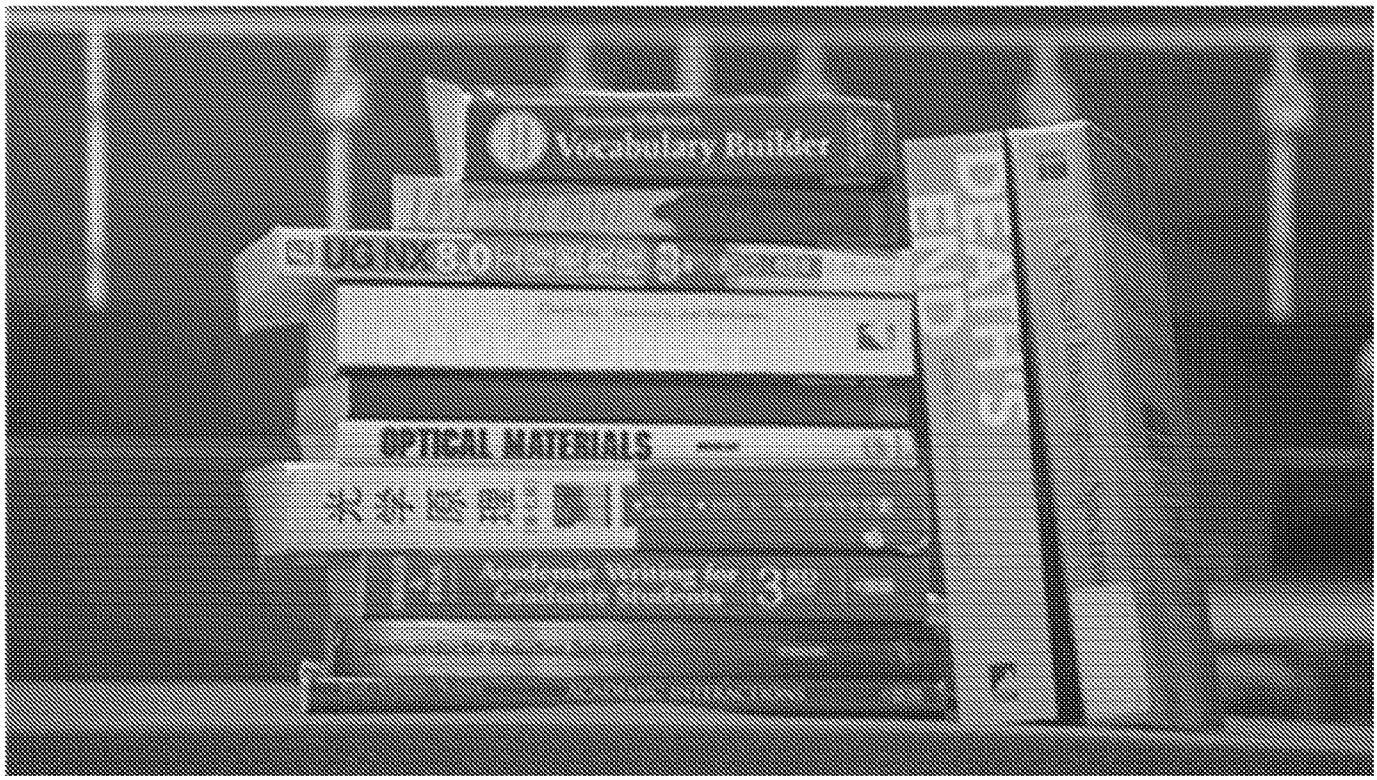
FIG. 7(A)



Example 1 (object distance 2.5m)

Out of Focus

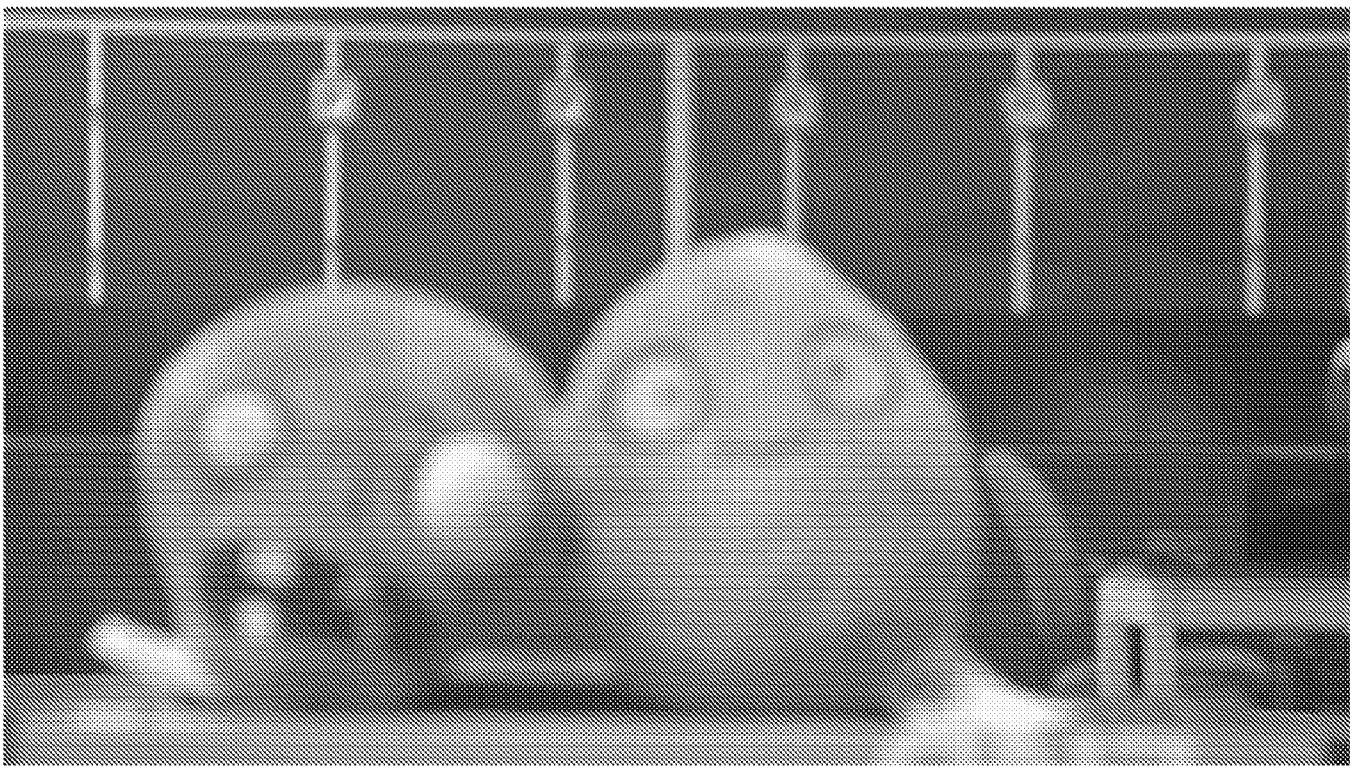
FIG. 7(B)



Example 1 (object distance 2.5m)

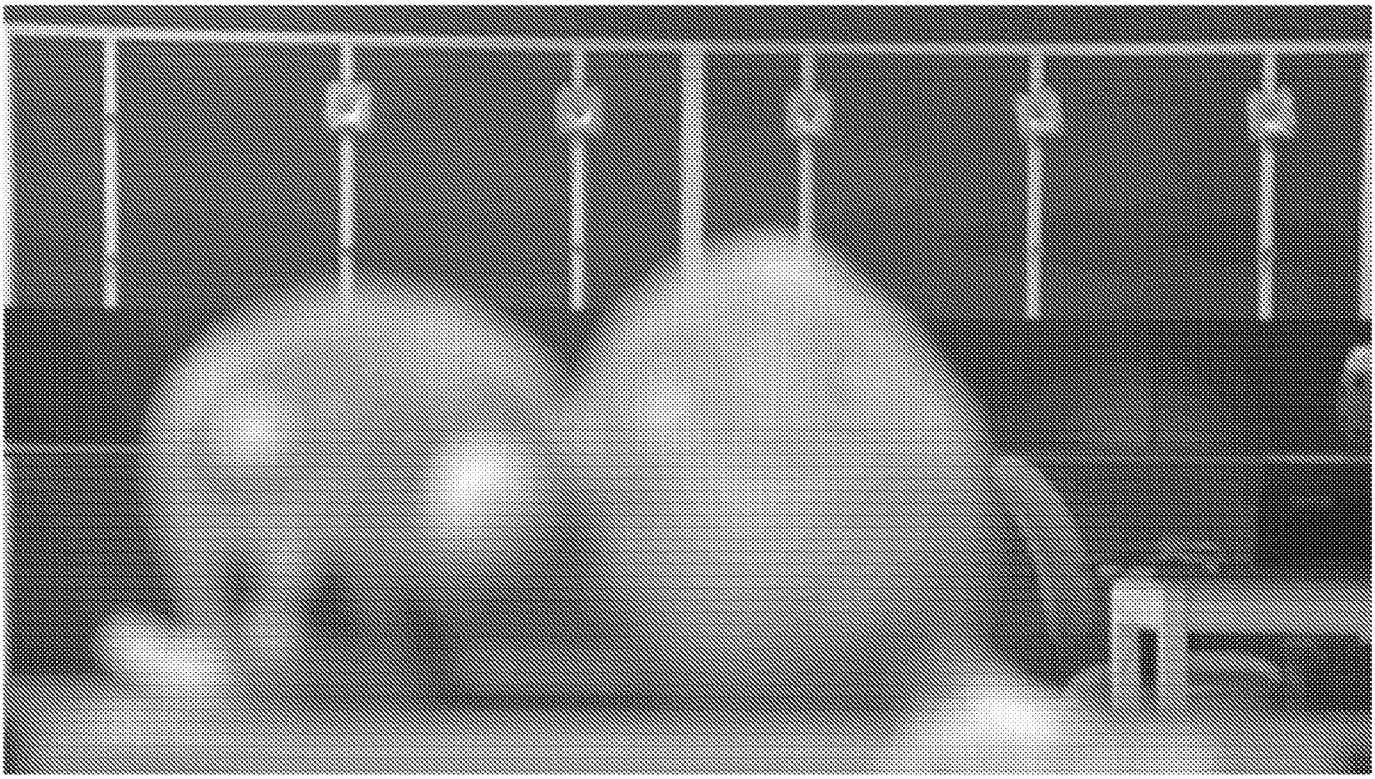
In Focus

FIG. 7(C)



Example 2 (object distance 1m)
Out of Focus

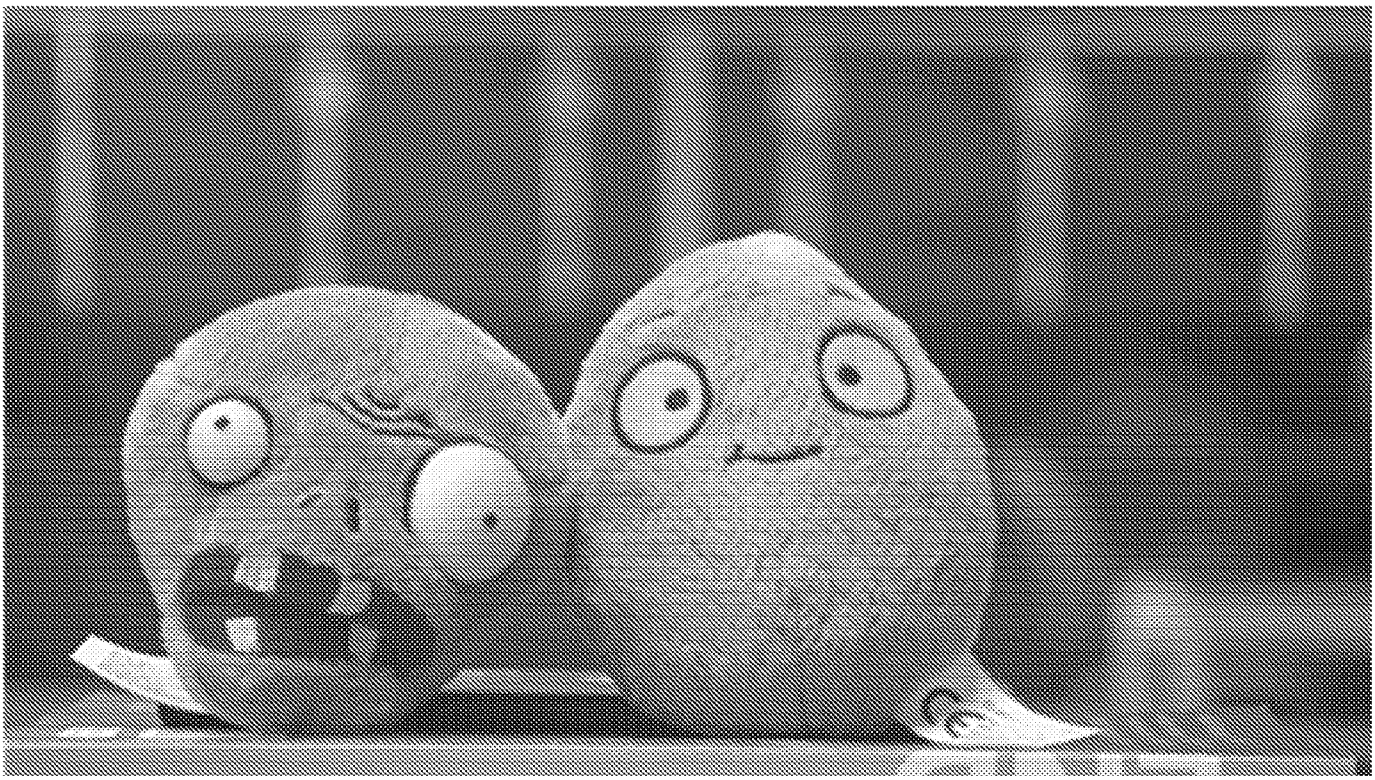
FIG. 8(A)



Example 2 (object distance 1m)

Out of Focus

FIG. 8(B)

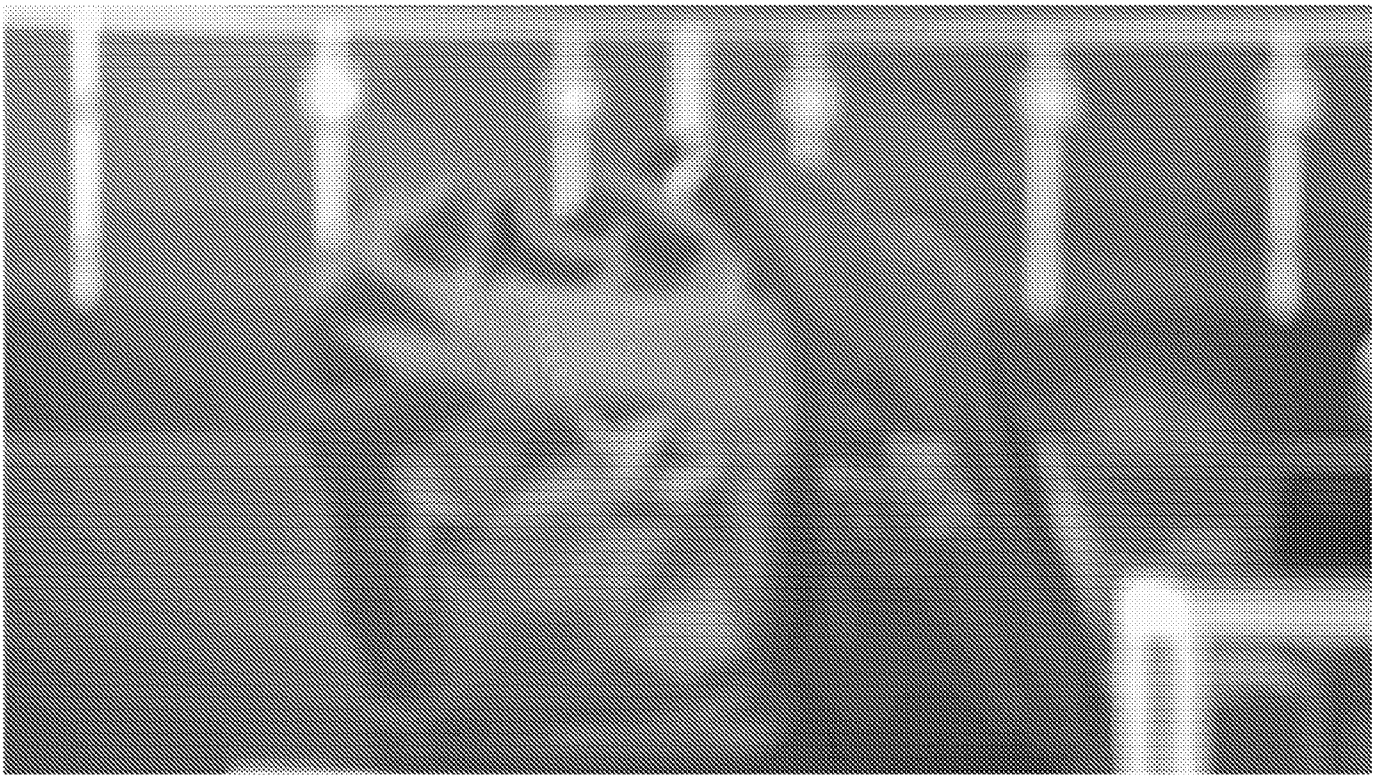


Example 2 (object distance 1m)
In Focus

FIG. 8(C)

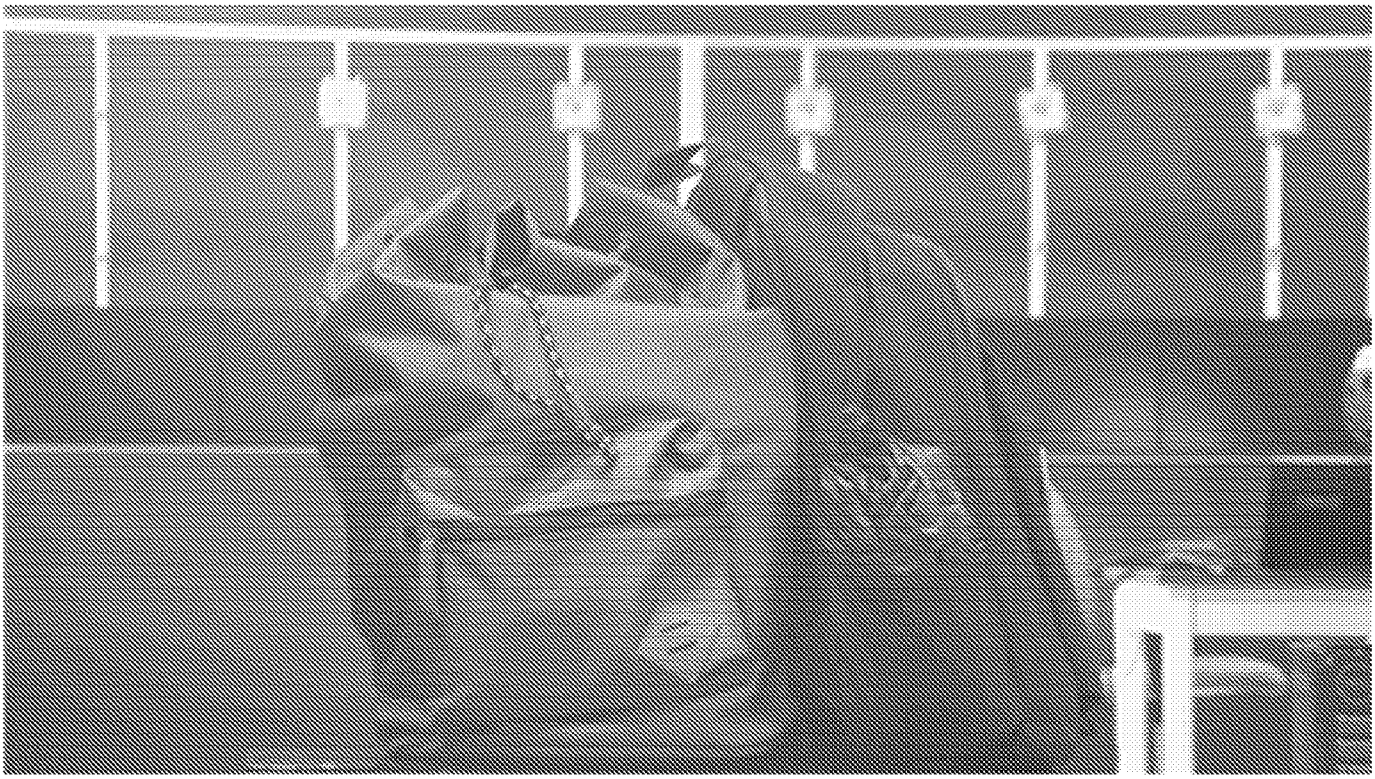
Example 3 (object distance 4.5m)
Out of Focus

FIG. 9(A)



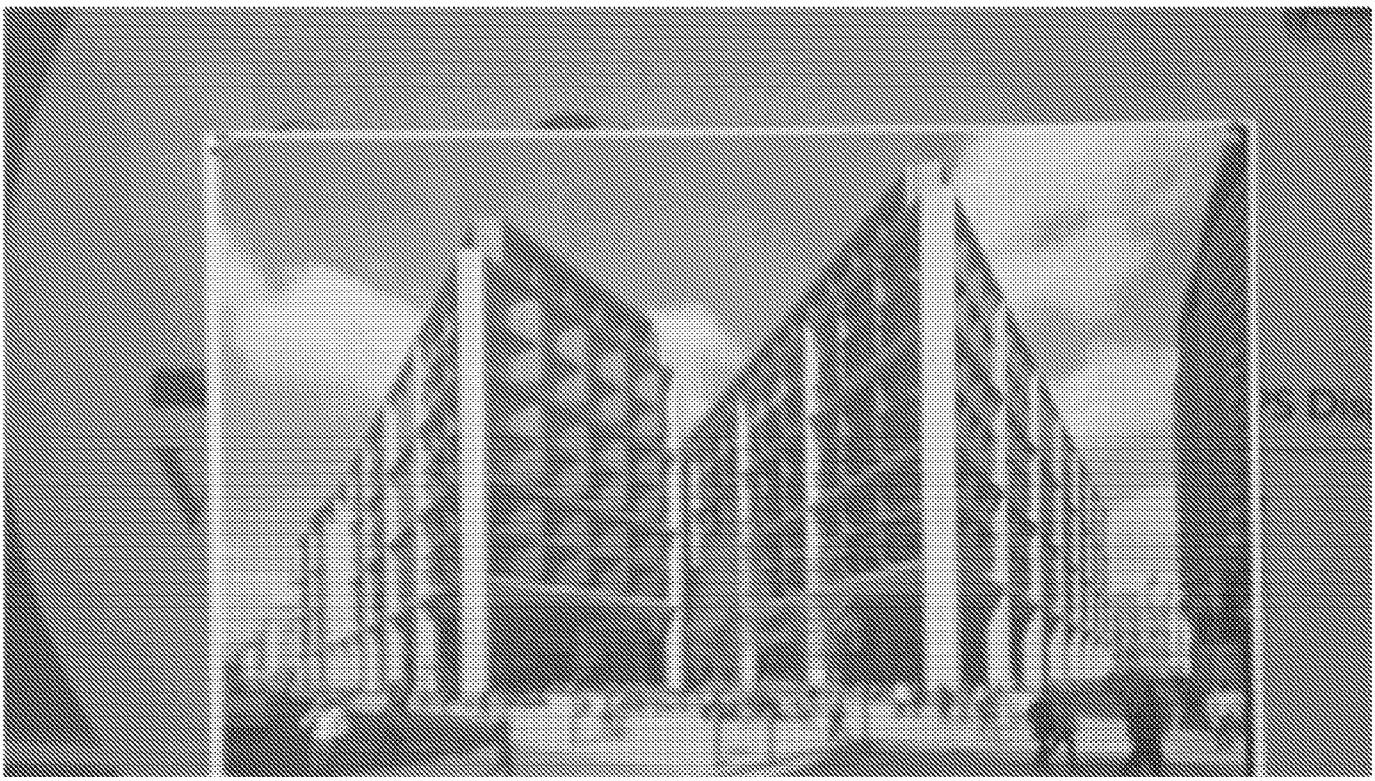
Example 3 (object distance 4.5m)
Out of Focus

FIG. 9(B)



Example 3 (object distance 4.5m)
In Focus

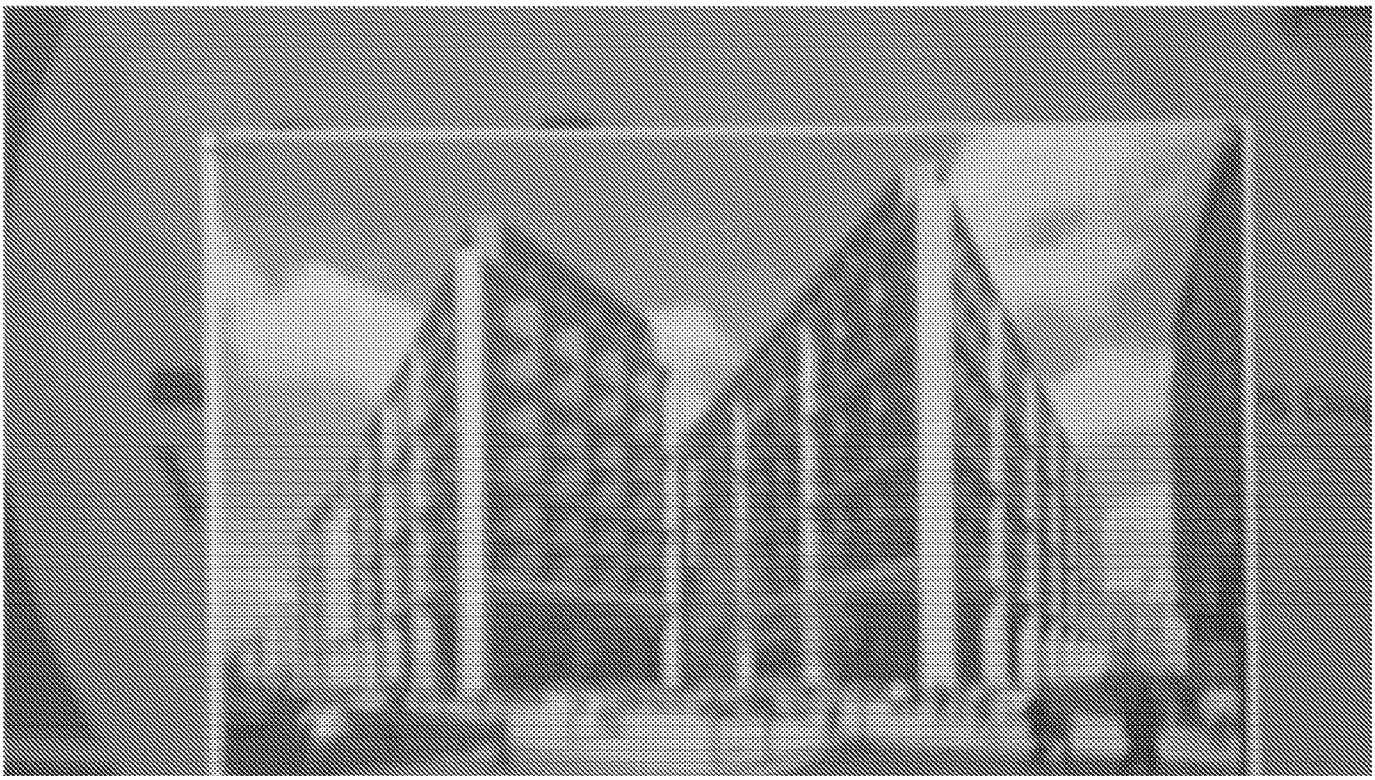
FIG. 9(C)



Example 4 (object distance 2.2m)

Out of Focus

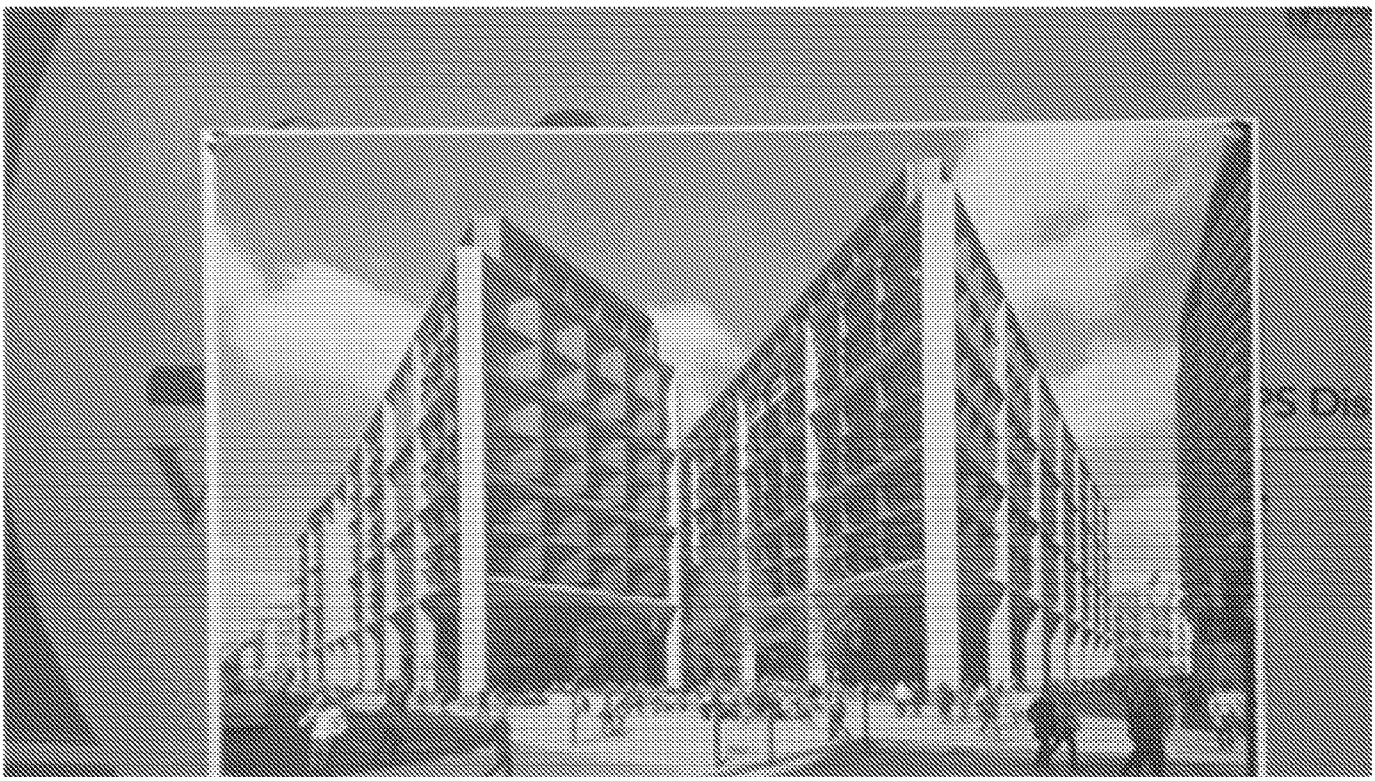
FIG. 10(A)



Example 4 (object distance 2.2m)

Out of Focus

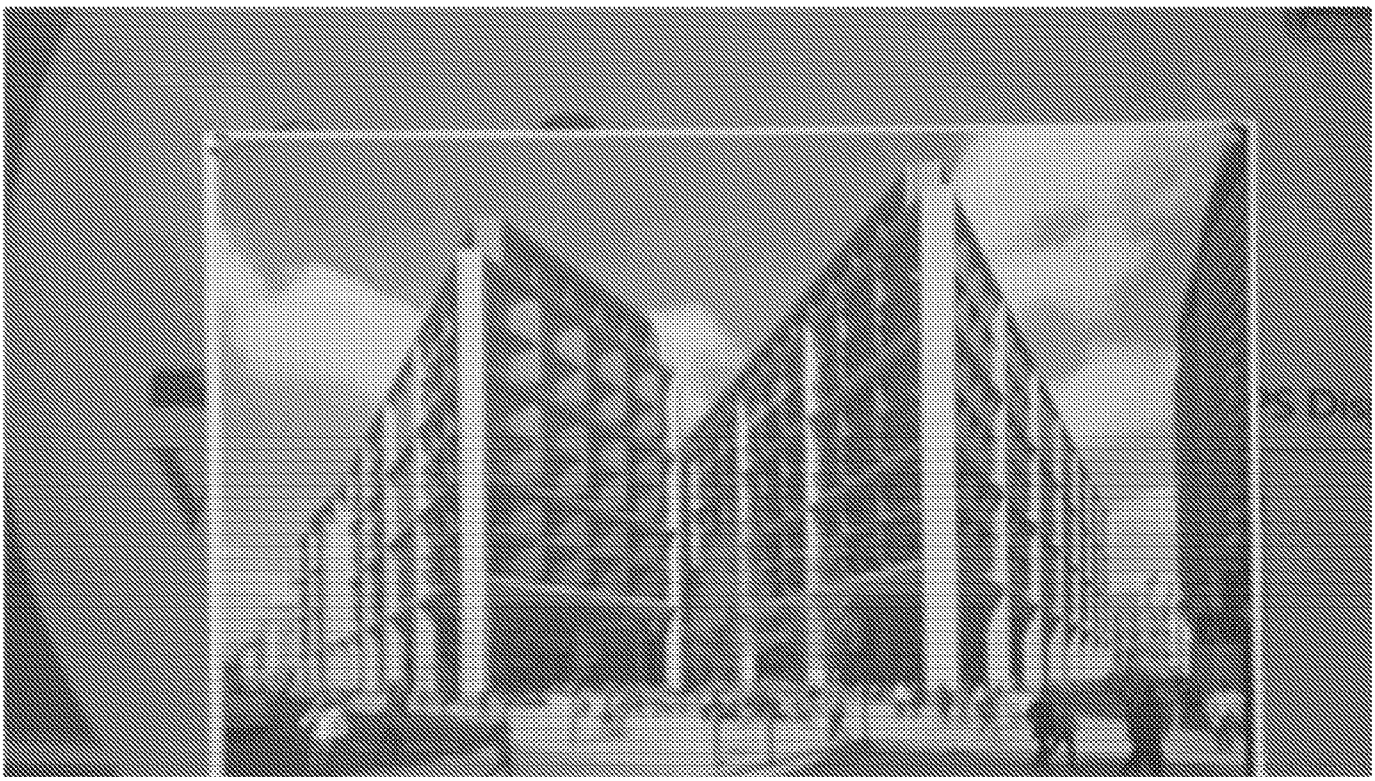
FIG. 10(B)



Example 4 (object distance 2.2m)

Out of Focus

FIG. 10(C)



Example 4 (object distance 2.2m)
Out of Focus

FIG. 10(D)

Example 4 (object distance 2.2m)
In Focus

FIG. 10(E)

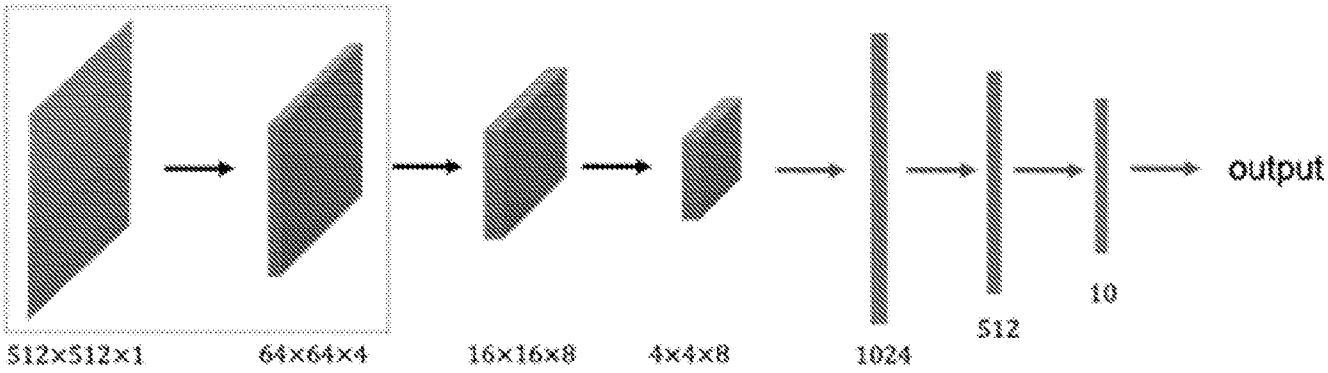
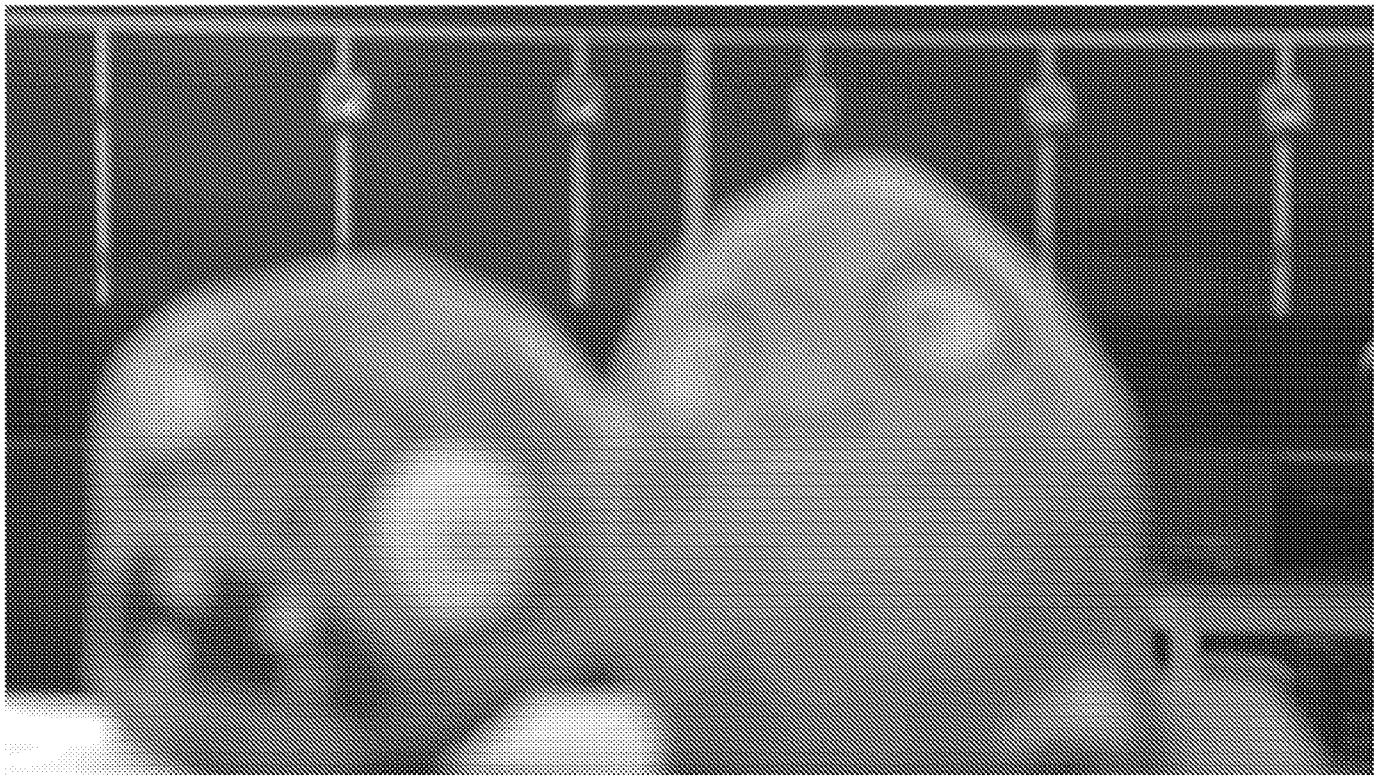
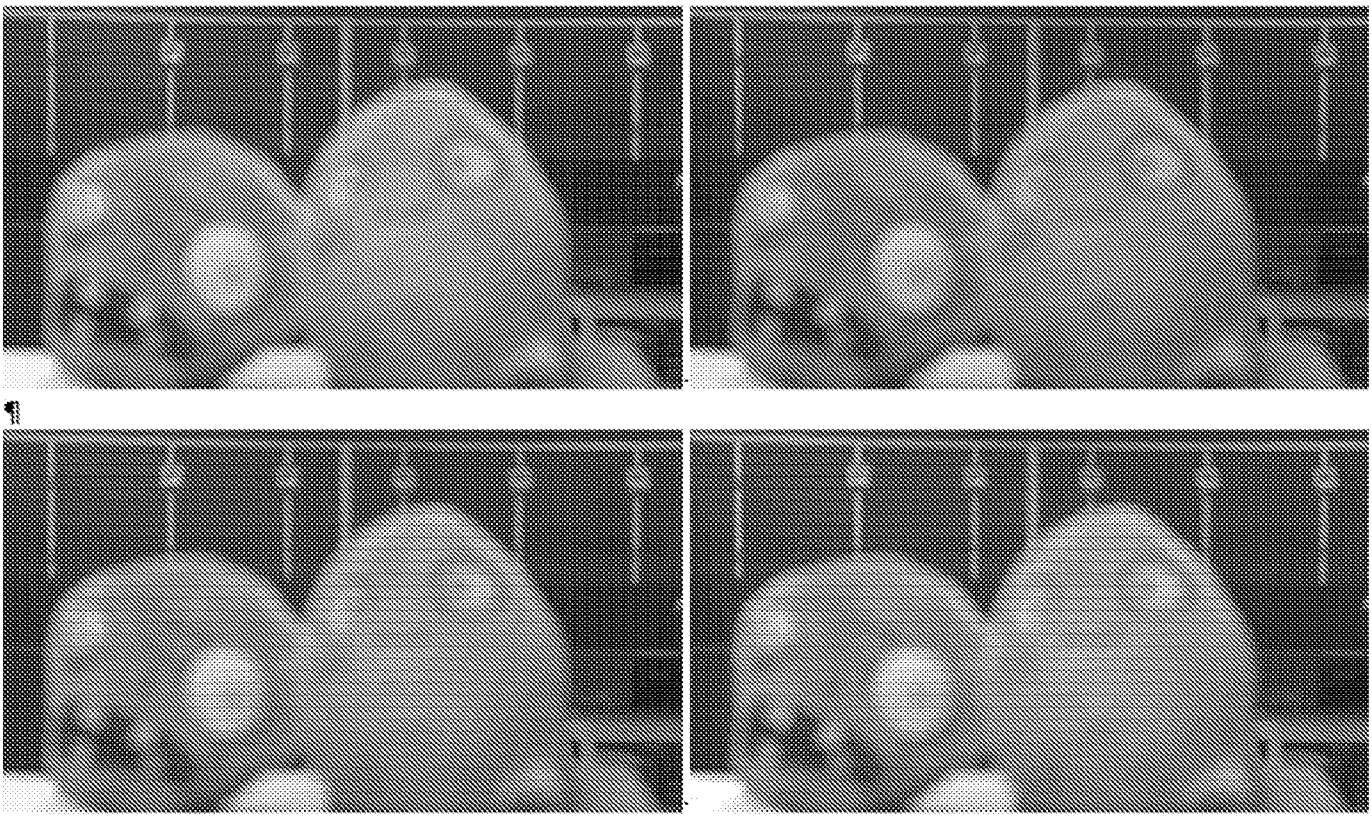


FIG 11



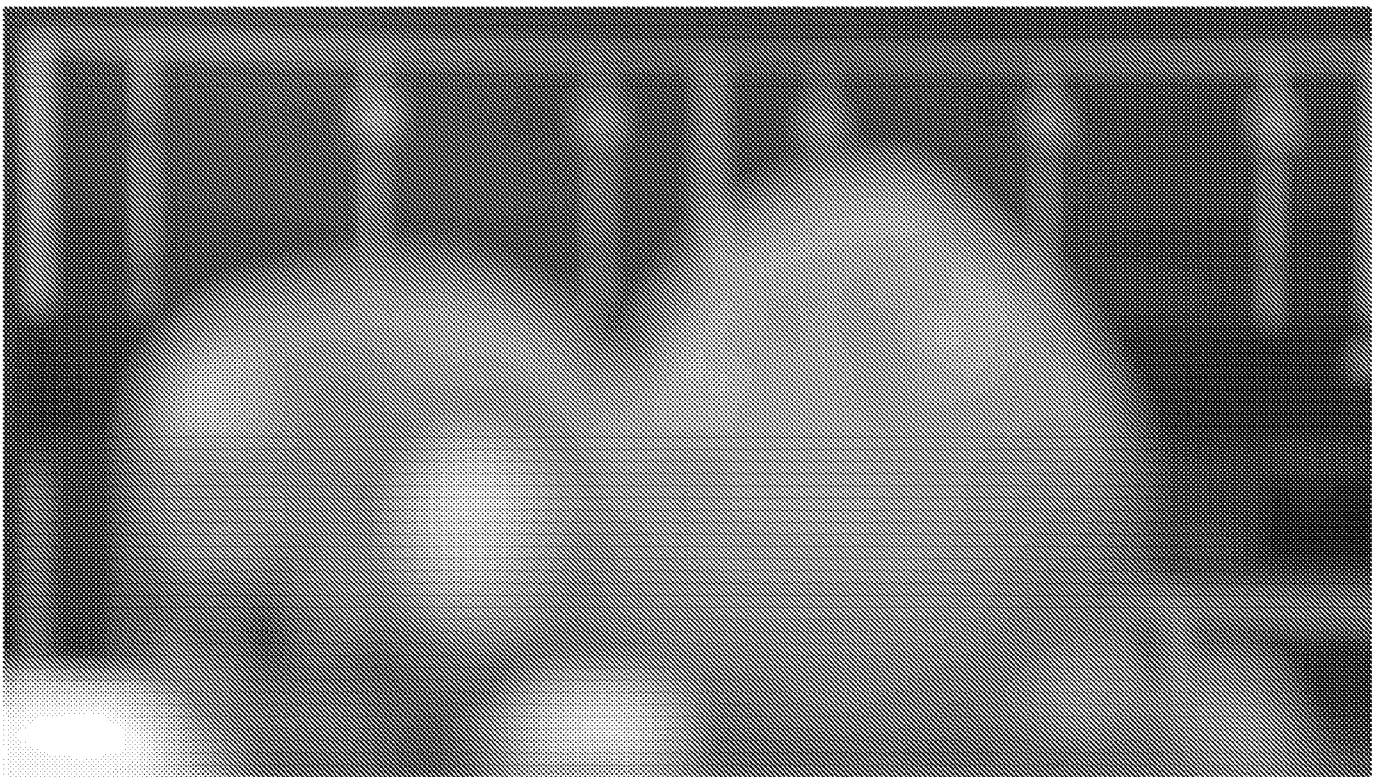
Example 5
Out of Focus

FIG. 12(A)



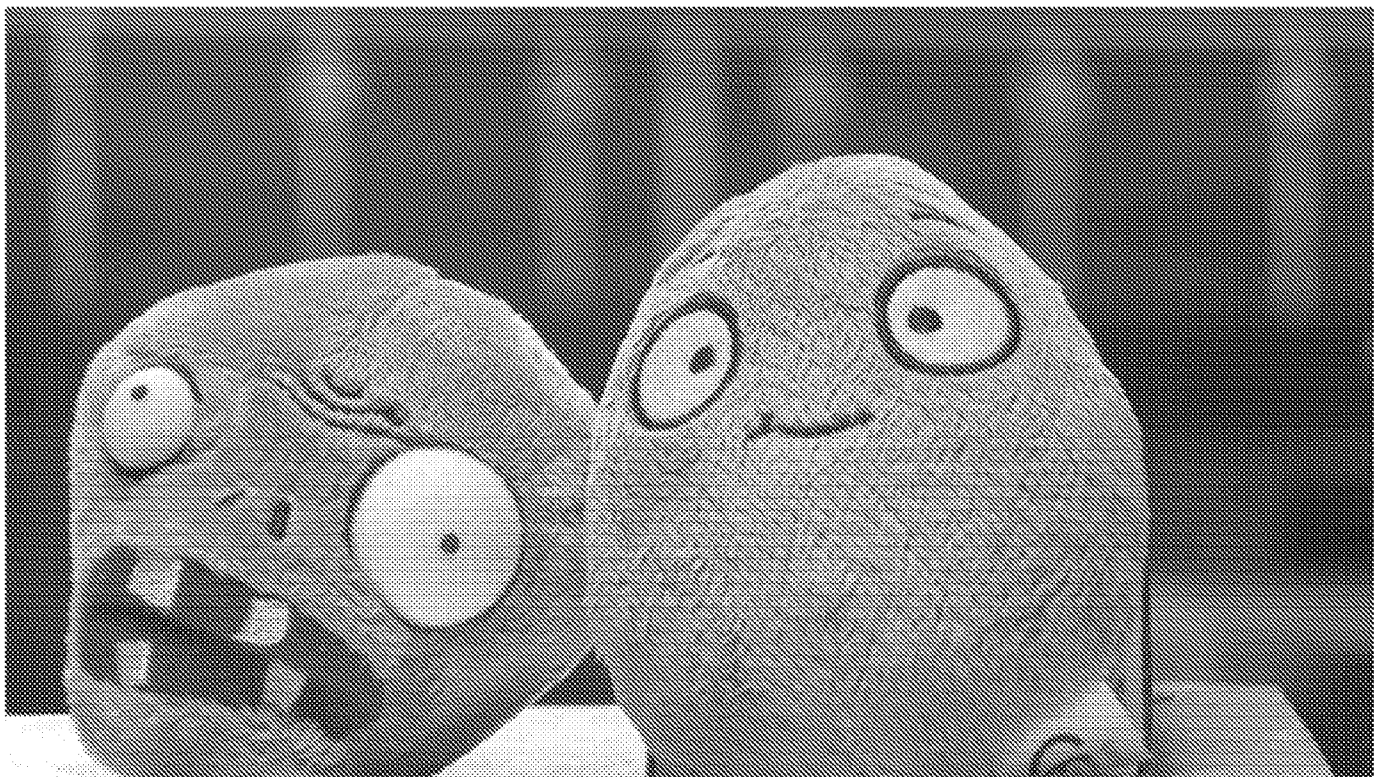
Example 5
Out of Focus

FIG. 12(B)



Example 5
Out of Focus

FIG. 12(C)



Example 5
In Focus

FIG. 12(D)

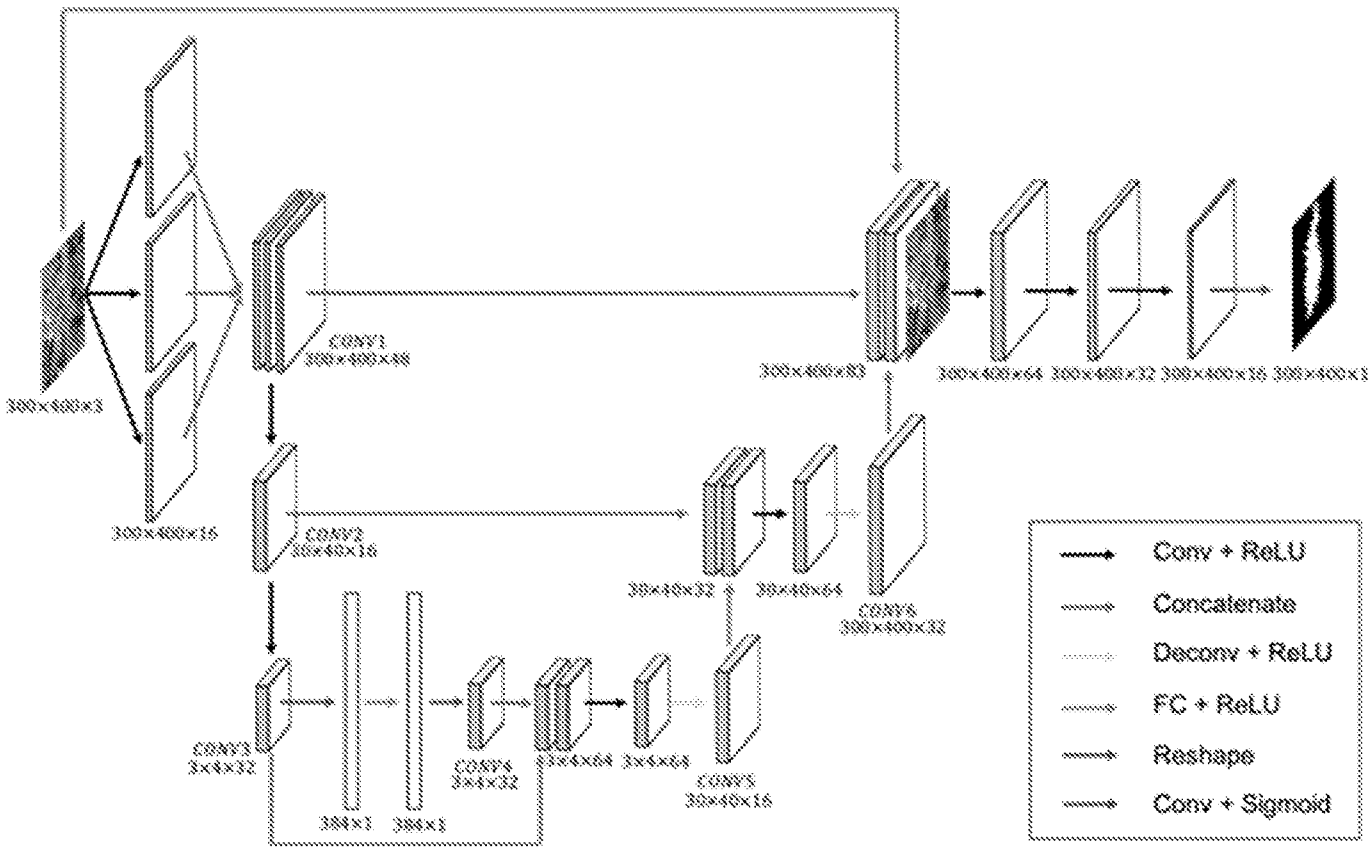
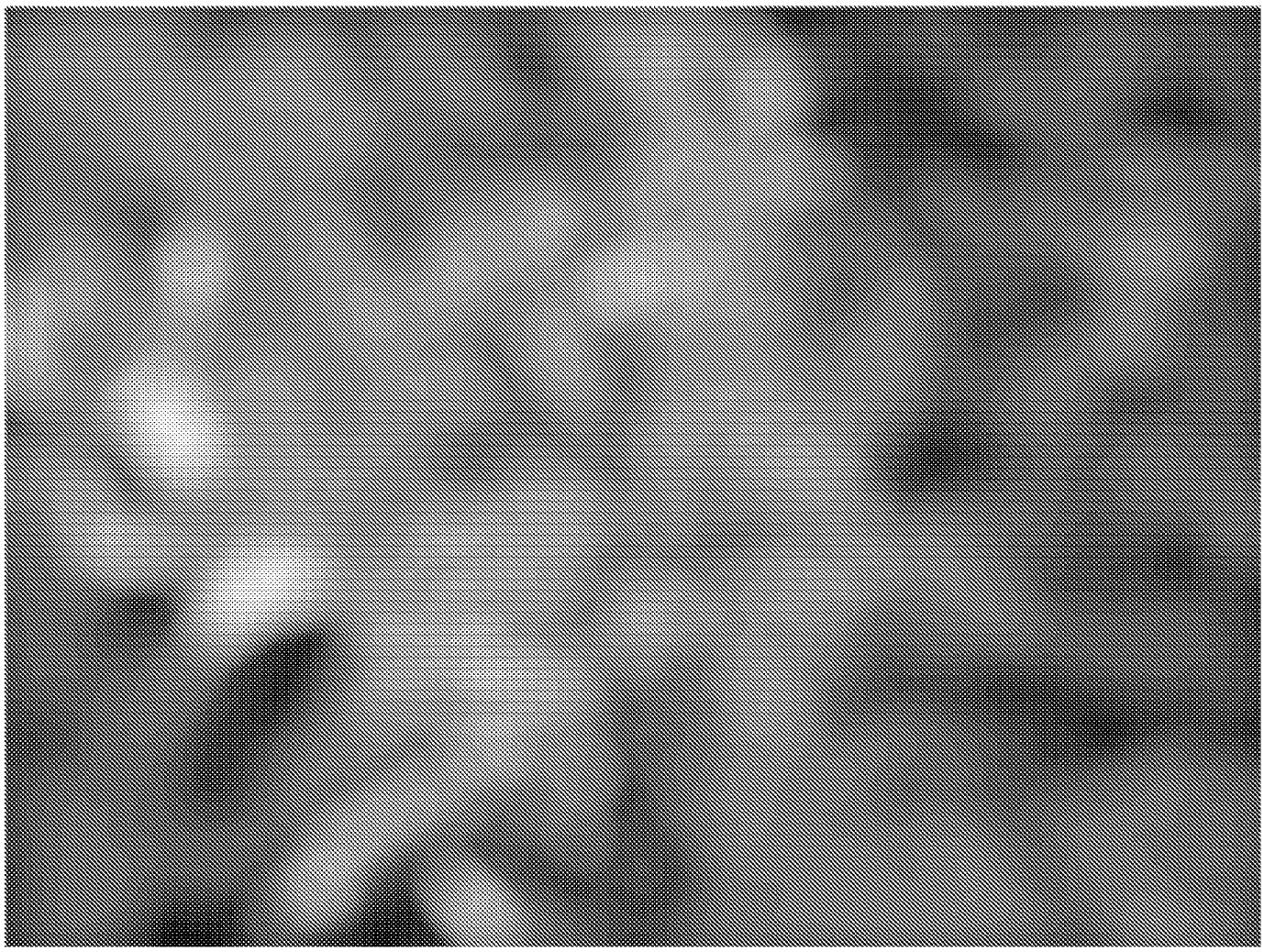

Conv + ReLU
Concatenate
Deconv + ReLU
FC + ReLU
Reshape
Conv + Sigmoid

FIG. 13

Original Image

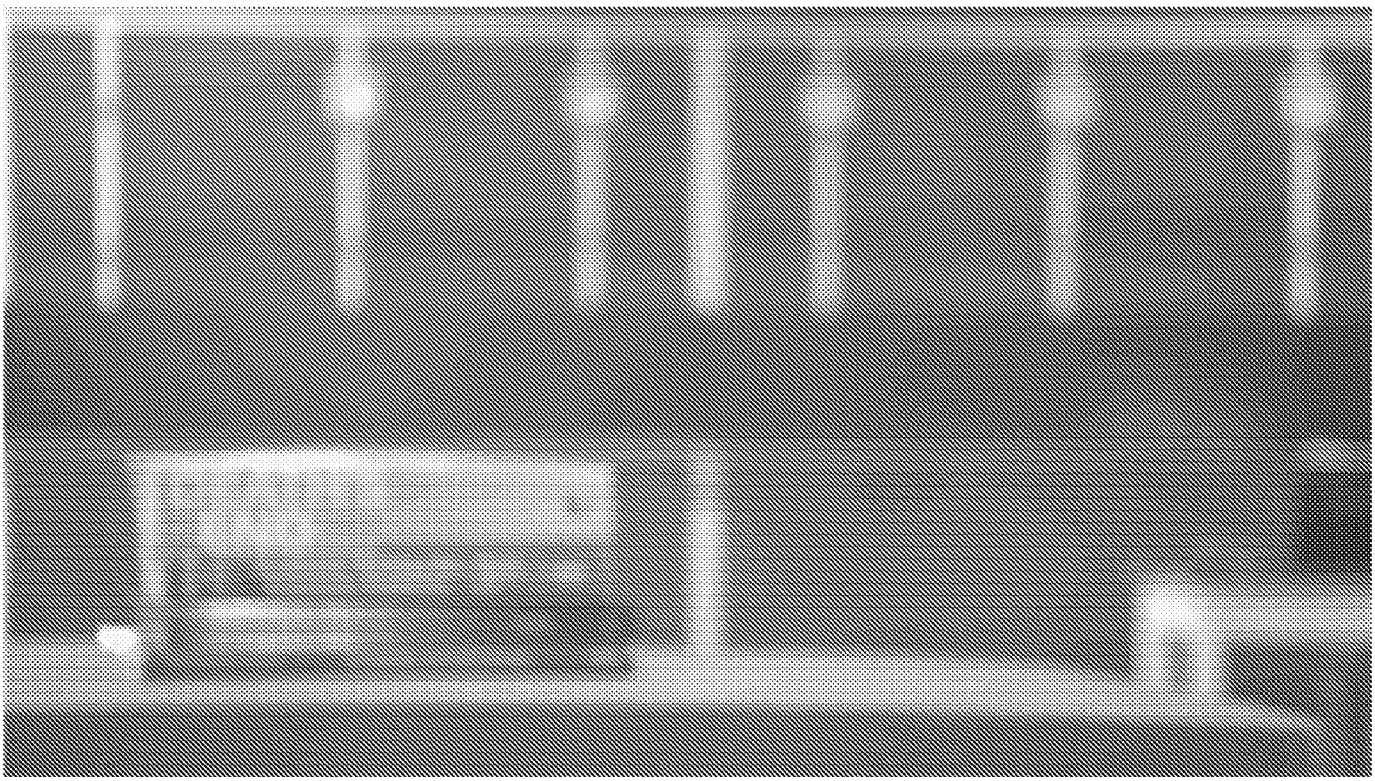
FIG. 14(A)



De-Focused Image

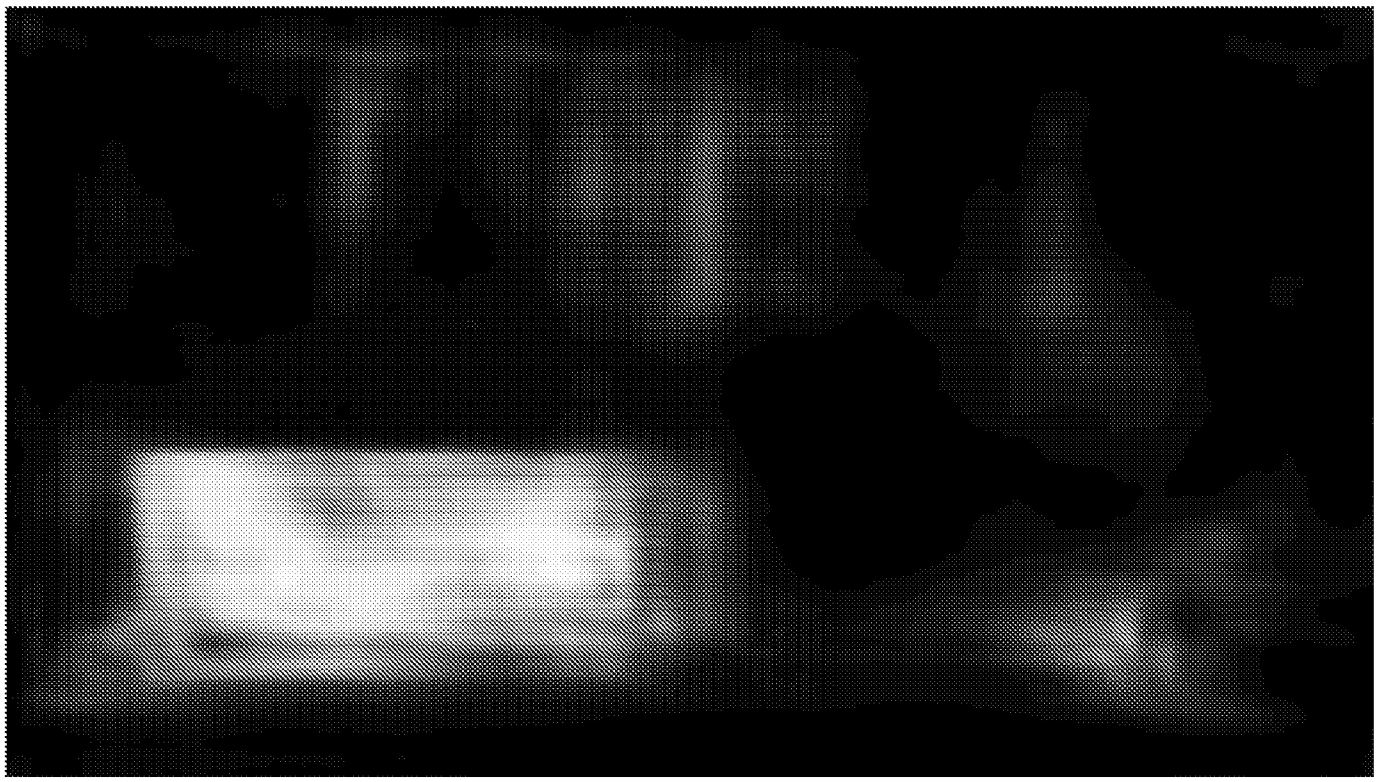
FIG. 14(B)

FIG. 14(C)



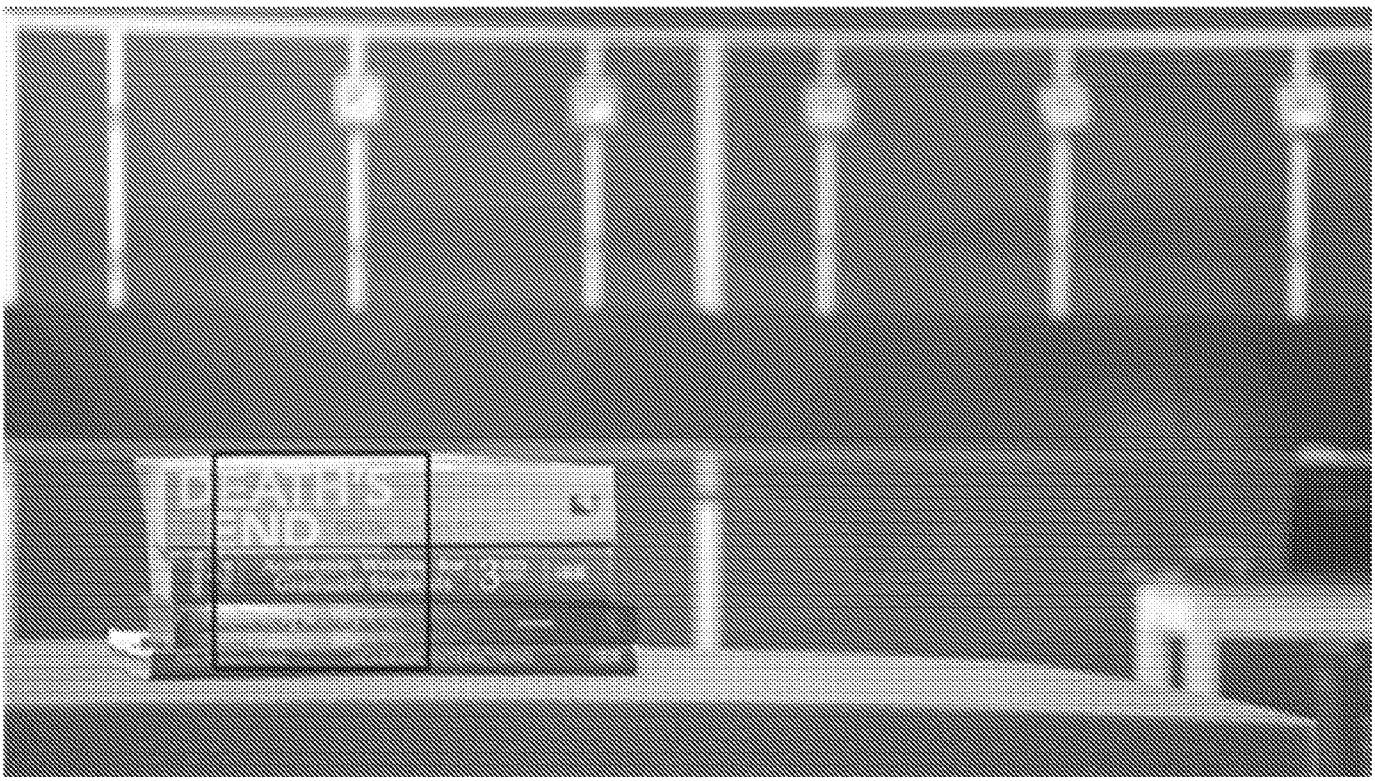
Example 6
Out of Focus

FIG. 15(A)



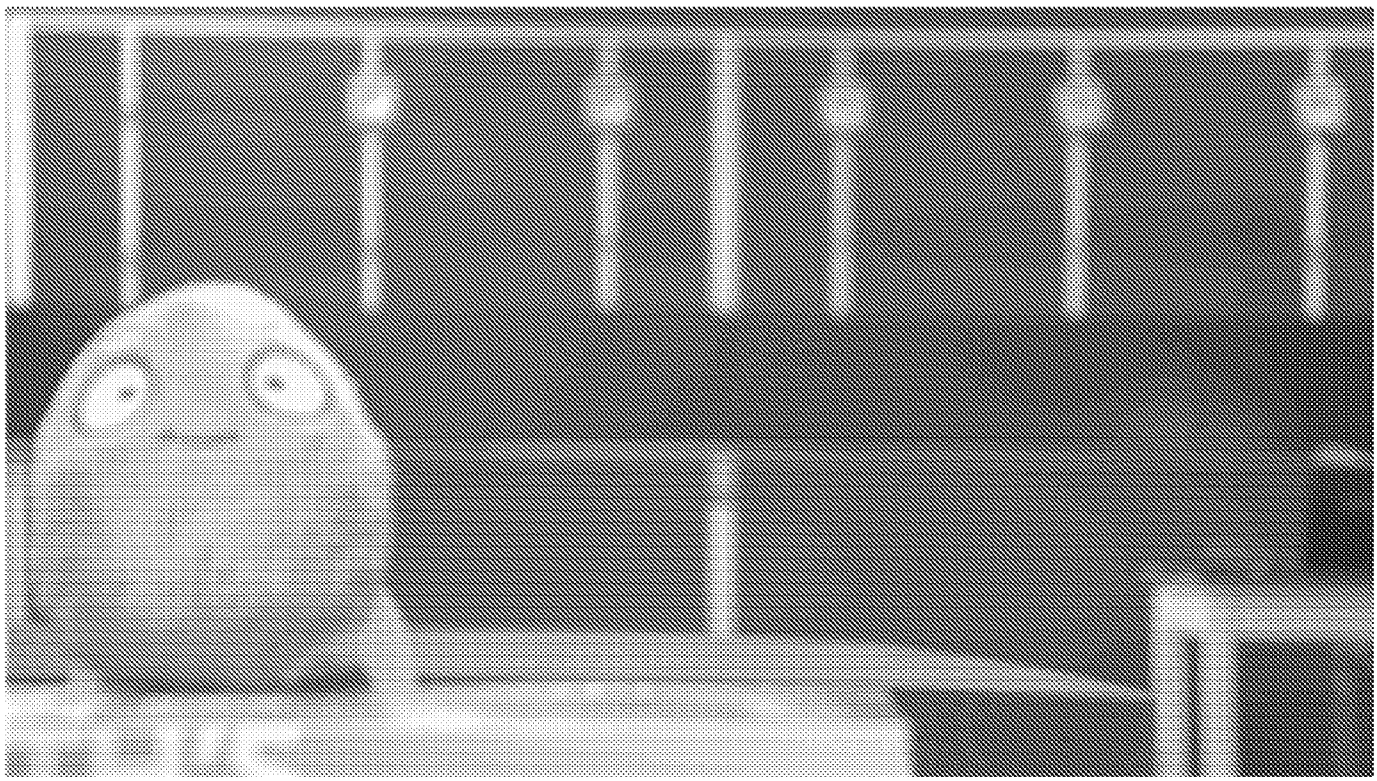
Example 6
Salience Map

FIG. 15(B)



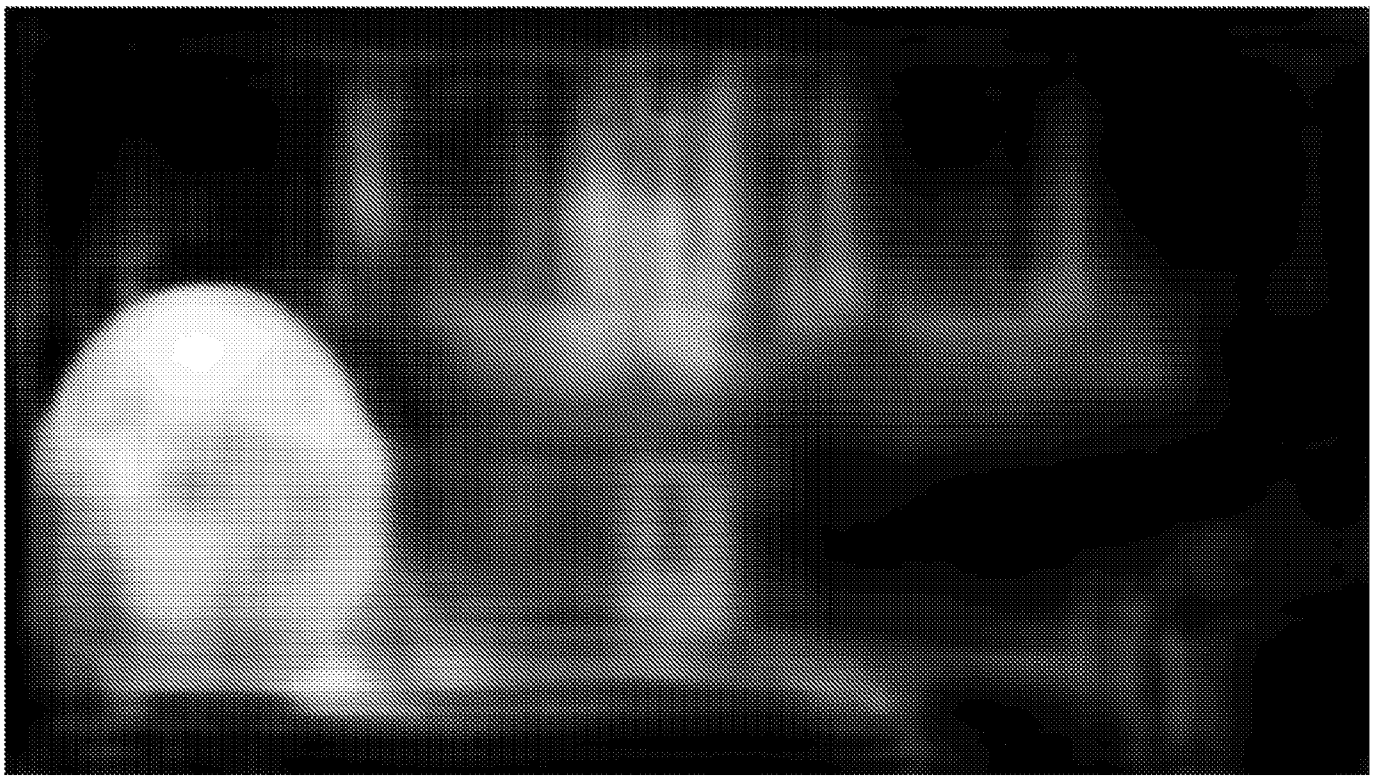
Example 6
Output Image

FIG. 15(C)



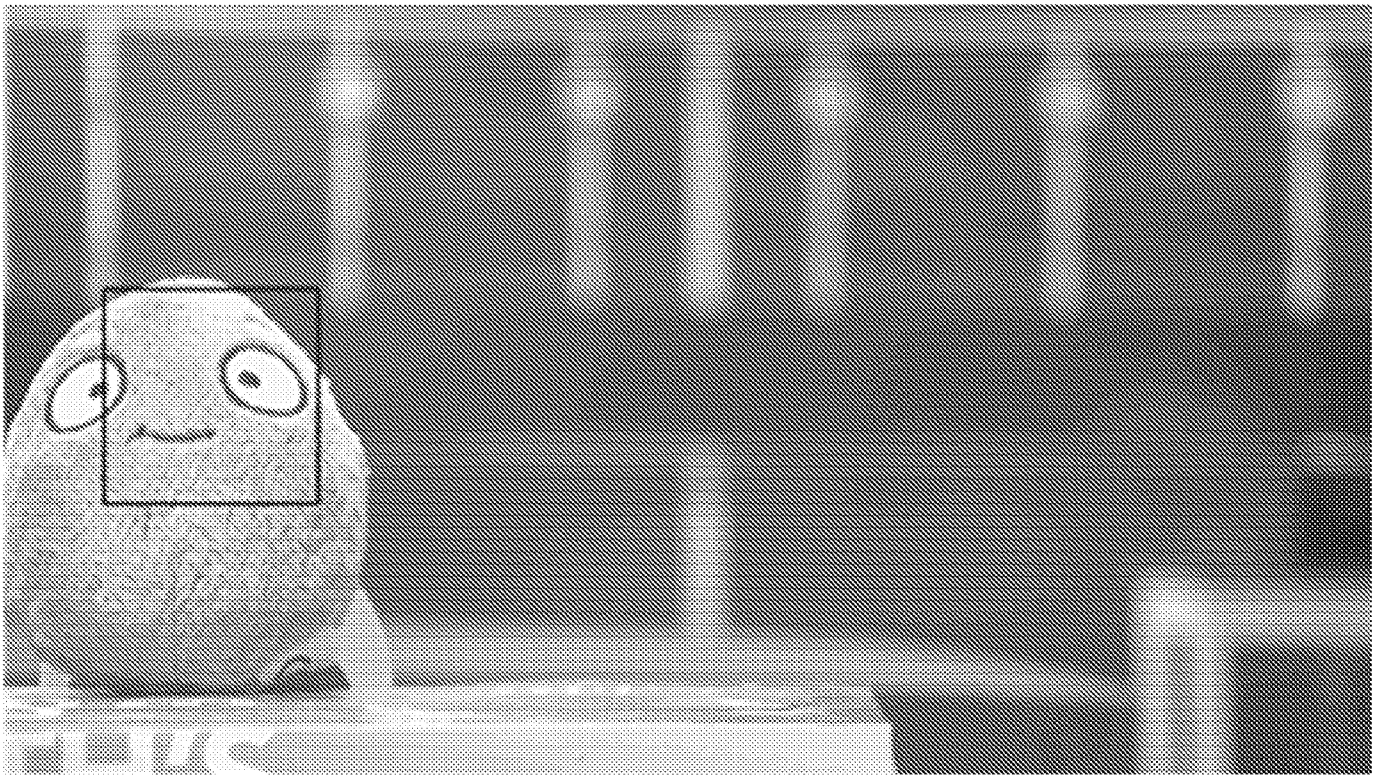
Example 7
Out of Focus

FIG. 16(A)



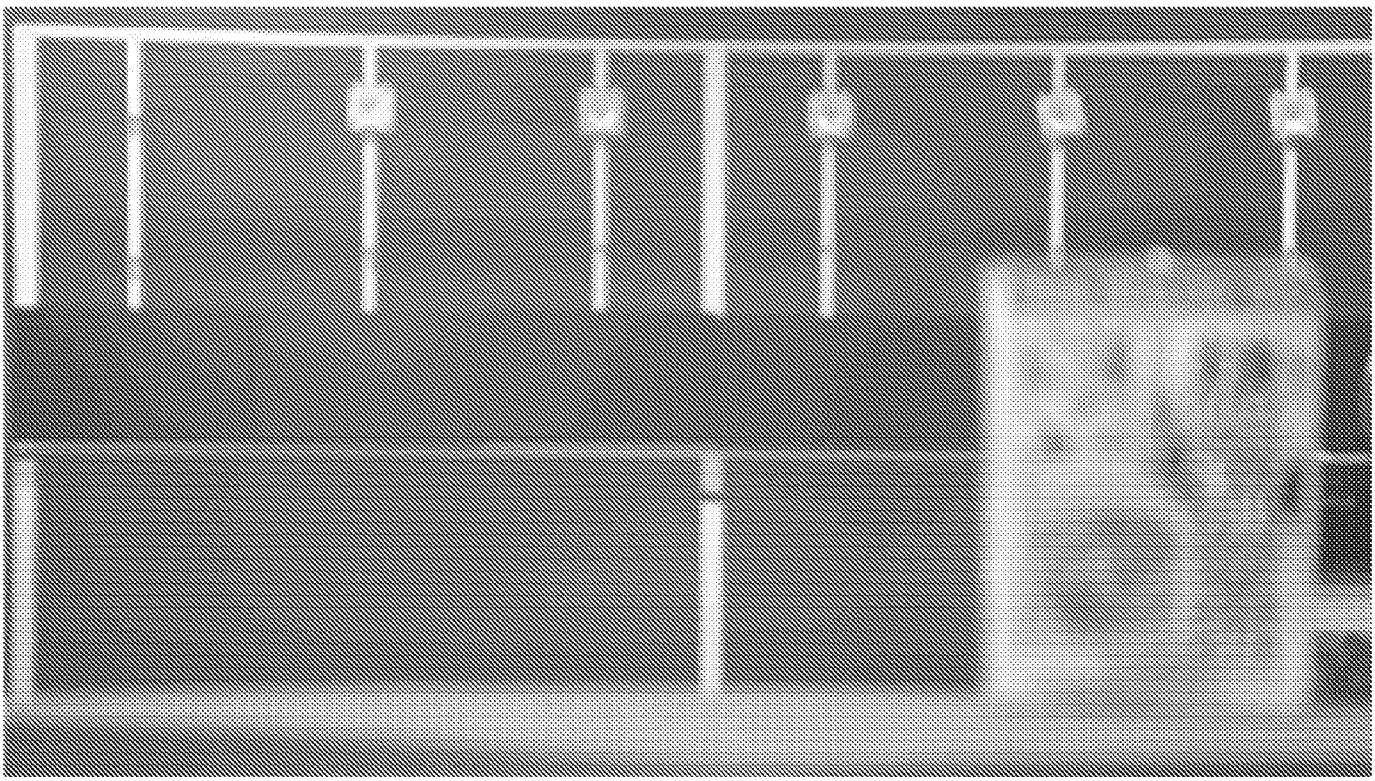
Example 7
Salience Map

FIG. 16(B)



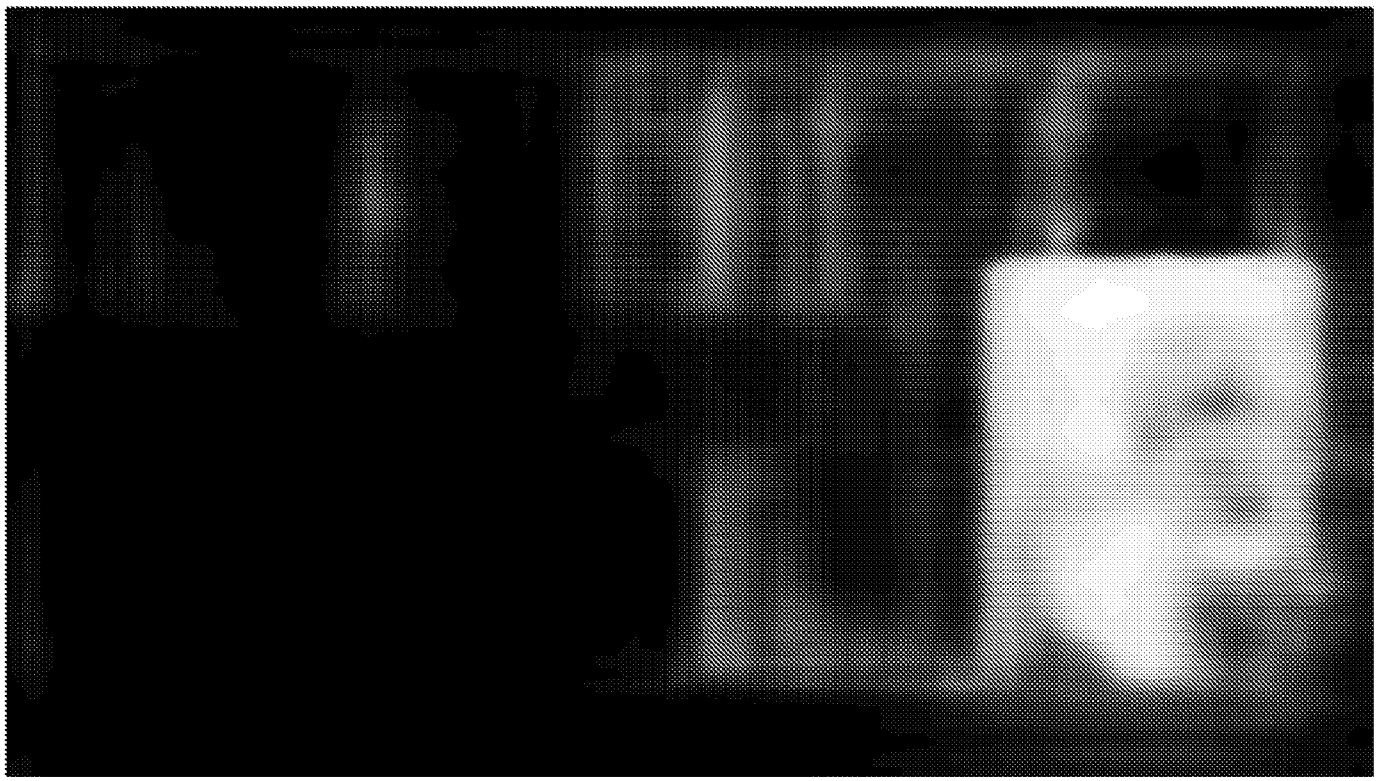
Example 7
Output Image

FIG. 16(C)



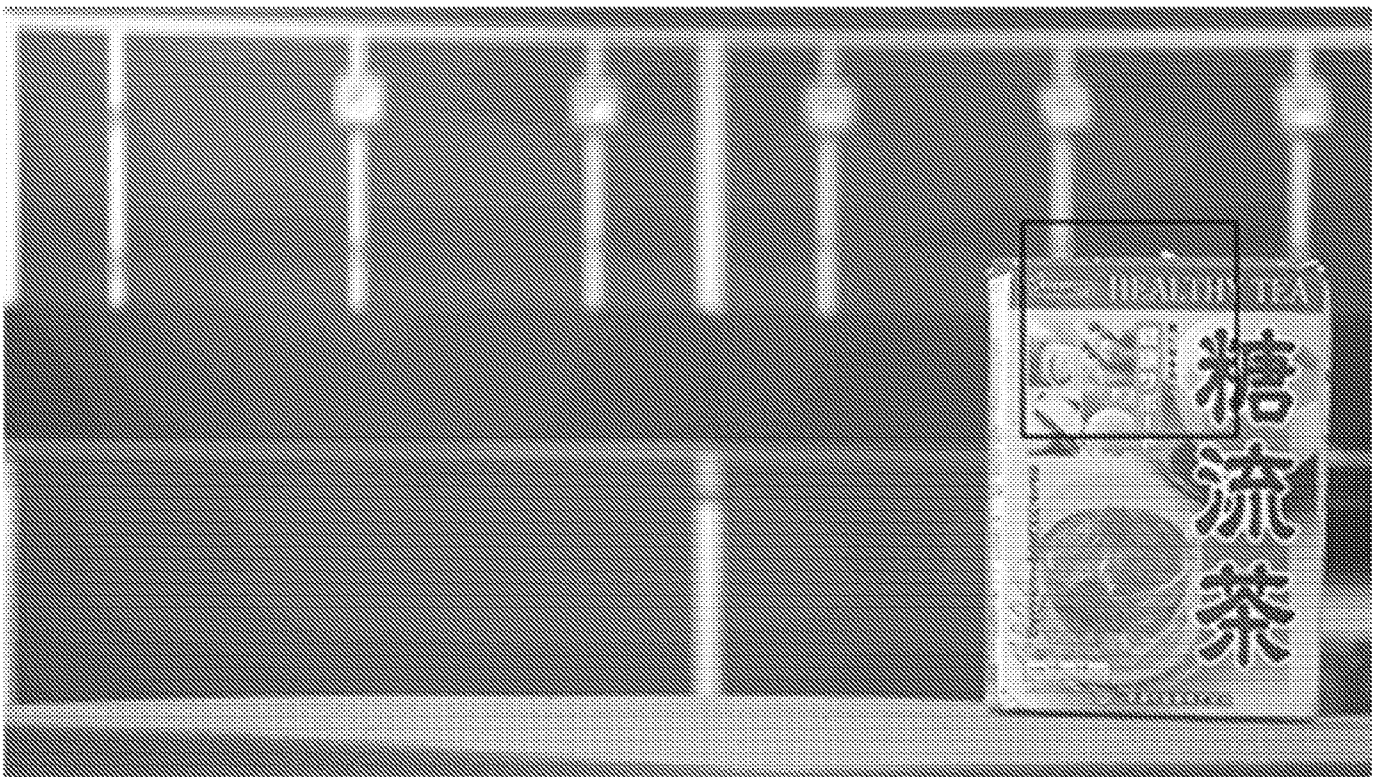
Example 8
Out of Focus

FIG. 17(A)



Example 8
Salience Map

FIG. 17(B)



Example 8
Output Image

FIG. 17(C)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/064755

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06T 5/20; G06T 7/00; G06T 7/20; H04N 5/232 (2020.01)
CPC - H04N 5/23254; G06T 5/20; G06T 2207/10024; H04N 5/23245; H04N 5/23267 (2020.01)

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 250/201.9 (keyword delimited)

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.
A	WEI et al. "Neural network control of focal position during time-lapse microscopy of cells." In: Scientific reports. 13 December 2017 (13.12.2017) Retrieved on 29 January 2020 (29.01.2020) from <https://www.biorxiv.org/content/10.1101/233940v1.full.pdf> entire document	1-10
A	US 2015/0116526 A1 (MENG et al) 30 April 2015 (30.04.2015) entire document	1-10
A	US 2011/0085028 A1 (SAMADANI et al) 14 April 2011 (14.04.2011) entire document	1-10
A	US 2017/0064204 A1 (DUKE UNIVERSITY) 02 March 2017 (02.03.2017) entire document	1-10
A	HE et al. "Modified fast climbing search auto-focus algorithm with adaptive step size searching technique for digital camera." In: IEEE transactions on Consumer Electronics. 03 April 2003 (03.04.2003) Retrieved on 29 January 2020 (29.01. 2020) from <https://pdfs.semanticscholar.org/c4d7/dcdcd2f541663bb8343b8ddd48765bb018db.pdf> entire document	1-10
A	YAO et al. "Evaluation of sharpness measures and search algorithms for the auto-focusing of high-magnification images." In: Proceedings Volume 6246, Visual Information Processing XV. 12 May 2006 (12.05.2006) Retrieved on 29 January 2020 (29.01.2020) from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.478.82&rep=rep1&type=pdf> entire document	1-10

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

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 January 2020	14 FEB 2020

Name and mailing address of the ISA/US	Authorized officer
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)