

CLAIMS

1. A human face and hand association detecting method, comprising:
acquiring a first image, the first image being an image of a human object;
performing feature extraction on the first image to obtain first feature maps of multiple scales;
performing feature fusion on the first feature maps of multiple scales to obtain second feature maps of multiple scales, wherein the scales of the second feature maps are in one-to-one correspondence with the scales of the first feature maps; and
detecting associated human face position and hand position for the same human object in the first image, based on the obtained second feature maps of multiple scales,
wherein, detecting associated human face position and hand position for the same human object in the first image, based on the obtained second feature maps of multiple scales comprises:
performing a convolution process on at least one of the second feature maps of multiple scales, to obtain associated human face position and hand position for the same human object in the first image.
2. The method according to claim 1, wherein acquiring the first image comprises:
acquiring a second image, the second image being an image including at least one human object;
performing human body target detection on the second image to obtain a detection frame of any one human object of the at least one human object in the second image; and
determining an image region, which corresponds to the detection frame of the any one human object, in the second image as the first image of the any one human object.
3. The method according to claim 1 or 2, wherein performing feature extraction on the first image to obtain the first feature maps of multiple scales comprises:
adjusting the first image into a third image of a preset scale; and
inputting the third image into a Res-Net to obtain the first feature maps of multiple scales.
4. The method according to any one of claims 1 to 3, wherein performing feature fusion on the first feature maps of multiple scales to obtain the second feature maps of multiple scales comprises:
inputting the first feature maps of multiple scales into a feature pyramid network, and
performing the feature fusion through the feature pyramid network, to obtain the second

feature maps of multiple scales.

5. The method according to any one of claims 1 to 4, wherein, in a descending order of scales, a plurality of first feature maps are expressed as $\{C_1, \dots, C_n\}$, wherein n represents the number of the first feature maps, and n is an integer greater than 1; and

wherein performing feature fusion on the first feature maps of multiple scales to obtain the second feature maps of multiple scales comprises:

performing a convolution process on a first feature map C_n by using a first convolution kernel, to obtain a second feature map F_n corresponding to the first feature map C_n , wherein the scale of the first feature map C_n is the same as the scale of the second feature map F_n ;

performing a linear interpolation process on the second feature map F_n to obtain a first intermediate feature map F'_n corresponding to the second feature map F_n , wherein the scale of the first intermediate feature map F'_n is the same as the scale of a first feature map C_{n-1} ;

performing a convolution process on a first feature map C_i other than the first feature map C_n by using a second convolution kernel, to obtain a second intermediate feature map C'_i corresponding to the first feature map C_i , wherein the scale of the second intermediate feature map C'_i is the same as the scale of a first intermediate feature map F'_{i+1} , wherein i is an integer variable greater than or equal to 1 and less than n ; and

obtaining a second feature map F_i other than the second feature map F_n by using the second intermediate feature map C'_i and the corresponding first intermediate feature map F'_{i+1} , wherein the first intermediate feature map F'_{i+1} is obtained by making the corresponding second feature map F_{i+1} subjected to a linear interpolation.

6. The method according to claim 5, wherein obtaining the second feature map F_i other than the second feature map F_n by using the second intermediate feature map C'_i and the corresponding first intermediate feature map F'_{i+1} comprises:

adding the second intermediate feature map C'_i and the corresponding first intermediate feature map F'_{i+1} to obtain the second feature map F_i .

7. The method according to any one of claims 1 to 6, wherein detecting the associated human face position and hand position for the same human object in the first image, based on the obtained second feature maps of multiple scales comprises:

performing a convolution process on the second feature map of the largest scale among the second feature maps of multiple scales, to obtain a mask map representing the human face

position and a mask map representing the hand position, respectively; and
determining a position region where the associated hand and human face are located in the first image, based on the mask map of the human face position and the mask map of the hand position.

8. The method according to any one of claims 1 to 7, wherein the method further comprises at least one of the following:

highlighting the associated hand and human face in the first image; and
allocating the same label to the associated human face position and hand position detected in the first image.

9. The method according to any one of claims 1 to 8, wherein the method is implemented through a neural network, and steps of training the neural network comprises:

acquiring a training image, wherein the training image is an image including a human object, and the training image has labeling information of actually associated human face position and hand position;

inputting the training image to the neural network, and predicting associated human face position and hand position for the same human object in the training image through the neural network; and

determining network loss based on the predicted associated human face position and hand position and the labeling information, and adjusting network parameters of the neural network according to the network loss, till training requirements are met.

10. A human face and hand association detecting device, comprising:

an acquisition module, configured to acquire a first image, the first image being an image of a human object;

a feature extraction module, configured to perform feature extraction on the first image to obtain first feature maps of multiple scales;

a fusion module, configured to perform feature fusion on the first feature maps of multiple scales to obtain second feature maps of multiple scales, wherein the scales of the second feature maps are in one-to-one correspondence with the scales of the first feature maps; and

a detection module, configured to detect associated human face position and hand position for the same human object in the first image, based on the obtained second feature maps of multiple scales,

wherein, the detection module is configured to perform a convolution process on at least one of the second feature maps of multiple scales, to obtain associated human face position and

hand position for the same human object in the first image.

11. The device according to claim 10, wherein the detection module is further configured to: perform a convolution process on the second feature map of the largest scale among the second feature maps of multiple scales, to obtain a mask map representing the human face position and a mask map representing the hand position, respectively; and determine a position region where the associated hand and human face are located in the first image, based on the mask map of the human face position and the mask map of the hand position.

12. The device according to claim 10 or 11, wherein the device comprises a neural network, and the feature extraction module, the fusion module and the detection module apply the neural network; and

the device further comprises a training module configured to train the neural network, wherein the steps of training the neural network comprises:

acquiring a training image, wherein the training image is an image including a human object, and the training image has the labeling information of actually associated human face position and hand position;

inputting the training image to the neural network, and predicting associated human face position and hand position for the same human object in the training image through the neural network; and

determining network loss based on the predicted associated human face position and hand position and the labeling information, and adjusting network parameters of the neural network according to the network loss, till training requirements are met.

13. An electronic apparatus, comprising:

a processor; and

a memory for storing processor-executable instructions;

wherein the processor is configured to invoke the instructions stored in the memory, to execute the method according to any one of claims 1 to 9.

14. A computer-readable storage medium having computer program instructions stored thereon, wherein the computer program instructions, when executed by a processor, implement the method according to any one of claims 1 to 9.

15. A computer program, wherein the computer program comprises computer-readable codes, and when the computer-readable codes run in an electronic apparatus, a processor in the electronic apparatus executes the method according to any one of claims 1 to 9.