

METHOD AND SYSTEM FOR FACE IN VIVO DETECTION

FIELD OF THE INVENTION

The present invention relates to a method and a system for face in vivo detection and in particular, but not exclusively, for face in vivo detection based on an illumination component.

BACKGROUND

The following discussion of the background to the invention is intended to facilitate an understanding of the present invention. However, it should be appreciated that the discussion is not an acknowledgment or admission that any of the material referred to was published, known or part of the common general knowledge in any jurisdiction as at the priority date of the application.

In recent years, biometric identification technology has made great progress and some common biometric features used for identification include human face, fingerprint, and iris. Biometric information for personal identification has been widely used in the world, through which real users and fake users can be distinguished. However, the accuracy of biometric identification can be compromised by various means, such as the use of forged images of human face, fingerprint, iris, etc., during biometric verification. To address this issue, there exist biometric identification systems for in vivo detection to distinguish between biometric information submitted to such a system from a living individual and that of a non-living individual (such as forged images of a living individual), so as to prevent illegal forgers from stealing other people's biometric information for personal identification.

Biometric identification systems, in particular, face recognition identification systems have been widely used for personal identification, video surveillance and video information retrieval analysis in recent years due to the convenience of use, high acceptability and other advantages. However, from the research phase to the application phase of face recognition technology, security threats associated with this technology must be addressed to ensure reliability and security of face recognition systems. In general, forgery login to a face recognition system may adopt one or more of the

following methods: face images, face video clips and three-dimensional face model replica. Among them, a face image can be obtained more easily than a face video clip or a three-dimensional face model replica, and hence is more frequently used in forgery login to a face recognition system. It is thus necessary to design a face in vivo detection system that is protected against threats from forgery face image login for the purpose of practical application of a face recognition system. The technologies of face in vivo detection and face recognition are complementary; the advancement and maturity of the former technology would affect the practical applications of the latter technology.

Existing detection methods in the field of face in vivo detection to distinguish between an image of a face (or face image) and a real human face are typically as follows: 1) To estimate the three-dimensional depth information by motion. The difference between a real face and a face image is that a real human face is a three-dimensional object with depth information while a face image is a two-dimensional plane, differentiations between the two can be found by reconstructing a three-dimensional face with several photos taken during the head-turning action. The disadvantage of this method is that the three-dimensional facial reconstruction requires many photos of facial features for accurate tracking, which still needs major adjustment. In addition, the calculation based on three-dimensional facial reconstruction method is very complicated; hence it is not possible to achieve real-time application based on this. 2) To distinguish between the two by analyzing the high-frequency component ratio of a face image and a real human face. The basic assumption of this method is that the face image imaging compared to real face imaging loses high frequency information. This method can effectively detect low-resolution face images, but does not apply to high-resolution photos. 3) To extract features from face images and design classifiers to distinguish between a face image and a real face. This method does not take into account three-dimensional geometric information in real faces, making it difficult to achieve ideal distinction accuracy. 4) Judgment based on interaction. The system sends all kinds of movement instructions to users randomly (such as turning head, nodding, opening mouth, blink, etc.), users perform the corresponding actions, and subsequently the system distinguishes between real faces and face images by analyzing these actions. This method requires the analysis of all kinds of actions, requires many complex algorithms for the analysis, and its verification accuracy and efficiency is far from satisfaction. To analyze actions like the opening and closing of mouth and blinking of eyes requires precise tracking of the

feature points on human faces, which is a very big challenge. In addition, the method requires users to perform a variety of movements in strict accordance with instructions, which is not friendly to users.

Therefore, there is an urgent need to address the technical problems of existing methods for distinguishing between an image of a face and a real human face. The present invention seeks to provide a method and a system to overcome at least in part some of the aforementioned disadvantages.

SUMMARY OF THE INVENTION

Throughout this document, unless otherwise indicated to the contrary, the terms 'comprising', 'consisting of', and the like, are to be construed as non-exhaustive, or in other words, as meaning 'including, but not limited to'.

In accordance with a first aspect of the present invention, there is provided a method for face detection, comprising:

- capturing a facial movement by video and processing the video to obtain a plurality of facial images from a plurality of successive video frames;

- using the Lambertian model to render each facial image obtained and computing discrete cosine transform (DTC) to obtain an illumination component of each facial image;

- calculating the mean local variance for the illumination components of the facial images; and

- comparing the mean local variance with a predetermined threshold (T_h) to determine whether the facial image is an image of a real face.

Preferably, each of the facial images obtained is denoted as I_i , where i is a natural number.

Preferably, wherein the illumination component of each face image I_i , which, according to the Lambertian model, can be denoted as:

$$I_i(x,y)=R_i(x,y)L_i(x,y)$$

where R_i is the reflection component, representing the surface reflectance of the

facial image; L_i is the illumination component, representing the illumination and shadow of the facial image, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_i , to obtain:

$$\begin{aligned} f_i &= \log I_i \\ &= \log R_i + \log L_i \\ &\triangleq v_i + u_i \end{aligned}$$

where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i = \log R$, $u_i = \log L$, compute DCT for f_i , i.e.:

$$F_i(s, t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x, y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

$$\text{Where } C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{Others} \end{cases}, \quad C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{Others} \end{cases}$$

where N is the length and width of the image, and the high frequency coefficient of $F_i(s, t)$ is set at 0, i.e.:

$$F'_i(s, t) = \begin{cases} F_i(s, t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

where M is a parameter to be defined, which is generally set at 5, compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x, y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s, t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Take f'_i as the estimation of the illumination component, i.e.:

$$u_i(x, y) \approx f'_i(x, y) \quad (7)$$

Then the illumination component of the facial image can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x, y) = \exp(u_i(x, y))$$

Preferably, M has an empirical value of 5.

Preferably, the step of calculating the mean local variance for the illumination components of the face images obtained from T successive video frames comprises:

dividing the illumination component of each face image equally into aBb image blocks with aBb pixels contained in each block, and use B_{ij} to represent the No.j image block of the face image from the No.i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{ij})$$

where $\text{var}(B_{ij})$ is the variance of the pixel value of the image block B_{ij} .

Preferably, in the step of comparing the mean local variance (Avar) obtained with a predetermined threshold (Th), if the Avar value is greater than or equal to Th, the face image in the video is an image of a real face.

Preferably, if the Avar value is less than Th, the face image in the video is not an image of a real face.

Preferably, the threshold (Th) is set according to specific image quality, whereby a lower image resolution means a lower threshold Th.

In accordance with a second aspect of the present invention, there is described a system for face detection, comprising:

an acquisition unit configured to capture a facial movement by video and to process the video to obtain a plurality of facial images from a plurality of successive video frames;

a calculation unit configured to render each facial image obtained using the Lambertian model, compute discrete cosine transform (DTC) to obtain an illumination component of each facial image, and calculate the mean local variance for the illumination components of the facial images; and

a determination unit configured to compare the mean local variance with a predetermined threshold (Th) to determine whether the facial image is an image of a

real face.

Preferably, each of the facial images obtained is denoted as I_i , where i is a natural number.

Preferably, the illumination component of each face image I_i , which, according to the Lambertian model, can be denoted as:

$$I_i(x,y)=R_i(x,y)L_i(x,y)$$

where R_i is the reflection component, representing the surface reflectance of the facial image; L_i is the illumination component, representing the illumination and shadow of the facial image, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_i , to obtain:

$$\begin{aligned} f_i &= \log I_i \\ &= \log R_i + \log L_i \\ &\triangleq v_i + u_i \end{aligned}$$

where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i=\log R$, $u_i=\log L$, compute DCT for f_i , i.e.:

$$F_i(s,t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x,y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

Where $C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{others} \end{cases}$, $C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{others} \end{cases}$

where N is the length and width of the image, and the high frequency coefficient of $F_i(s,t)$ is set at 0, i.e.:

$$F'_i(s,t) = \begin{cases} F_i(s,t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

where M is a parameter to be defined, which is generally set at 5, compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x,y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s,t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Take f_i as the estimation of the illumination component, i.e.:

$$u_i(x, y) \approx f_i'(x, y) \quad (7)$$

Then the illumination component of the facial image can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x, y) = \exp(u_i(x, y))$$

Preferably, M has an empirical value of 5.

Preferably,

the step of calculating the mean local variance for the illumination components of the face images obtained from T successive video frames comprises:

dividing the illumination component of each face image equally into $a \times b$ image blocks with $a \times b$ pixels contained in each block, and use $B_{i,j}$ to represent the $No. j$ image block of the face image from the $No. i$ frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{i,j})$$

where $\text{var}(B_{i,j})$ is the variance of the pixel value of the image block $B_{i,j}$.

Preferably, in the step of comparing the mean local variance ($Avar$) obtained with a predetermined threshold (Th), if the $Avar$ value is greater than or equal to Th , the face image in the video is an image of a real face.

Preferably, if the $Avar$ value is less than Th , the face image in the video is not an image of a real face.

Preferably, the threshold (Th) is set according to specific image quality, whereby a lower image resolution means a lower threshold Th .

Other aspects and advantages of the invention will become apparent to those skilled in the art from a review of the ensuing description, which proceeds with reference to the following illustrative drawings of various embodiments of the invention.

DETAILED DESCRIPTION

Particular embodiments of the present invention will now be described. The terminology used herein is for the purpose of describing particular embodiments only and is not intended to limit the scope of the present invention. Additionally, unless 系 defined otherwise, all technical and scientific terms used herein have the same meanings as commonly understood by one of ordinary skill in the art to which this invention belongs.

The use of the singular forms `a\_`, `an\_`, and `the\_` include both singular and plural referents unless the context clearly indicates otherwise.

本 The use of `or\_`, `/\_` means `and/or\_` unless stated otherwise. Furthermore, the use of the terms `including\_` and `having\_` as well as other forms of those terms, such as `includes\_`, `included\_`, `has\_`, and `have\_` are not limiting.

In order to overcome at least in part some of the aforementioned disadvantages of existing methods for distinguishing between an image of a face (or a face image) and a 系 real face such as complex calculations, poor adaptability, low distinction accuracy, and low efficiency, the present invention provides a method and a system for face or face in vivo detection based on an illumination component which offers high accuracy, excellent real-time performance and improved user experience.

In order to achieve the above-mentioned technical objectives, the technical solution of 本 the present invention comprises a method for face or face in vivo detection based on an illumination component, wherein the method includes the following steps:

Step 1: capture a facial movement such as a human head movement in the form of a video and obtain a plurality of face images by cropping the video.

Step 2: use the Lambertian model to render each face image obtained from Step 1, then 本 compute discrete cosine transform (DCT) to obtain the illumination component of each face image. It would be appreciated that in other embodiments, it can be synchronized with surface shading algorithms, ray cast and the like.

Step 3: based on the results above in Step 2, calculate the mean local variance for the illumination component of each of the face images obtained from several successive

video frames.

Step 4: compare the mean local variance with a predetermined or predefined threshold to determine whether the face image is an image of a real face.

Step 1 above requires obtaining face images by cropping the human head movement video and the face images are denoted as I_j .

Step 2 above requires extracting the illumination component of each face image I_j , which, according to the Lambertian model, can be denoted as:

$$I_j(x,y) = R_j(x,y)L_j(x,y)$$

Where R_j is the reflection component, representing the surface reflectance of the image scene; L_j is the illumination component, representing the illumination and shadow of the image scene, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_j , to obtain:

$$\begin{aligned} f_j &= \log I_j \\ &= \log R_j + \log L_j \end{aligned}$$

$$f_j \triangleq v_j + u_j$$

Where f_j , v_j and u_j respectively represent the value of I , R and L over the log-domain, i.e. $v_j = \log R$, $u_j = \log L$,

Compute DCT (discrete cosine transform) for f_j , i.e.:

$$F_j(s,t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_j(x,y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

$$\text{Where } C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{Others} \end{cases}, \quad C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{Others} \end{cases}$$

Where N is the length and width of the image, and the high frequency coefficient of $F_j(s,t)$ is set at 0, i.e.:

$$F'_i(s, t) = \begin{cases} F_i(s, t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

Where M is a parameter to be defined, which is generally set at 5,

Compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x, y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s, t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

Take f'_i as the estimation of the illumination component, i.e.:

$$u_i(x, y) \approx f'_i(x, y)$$

Then the illumination component of the image domain can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x, y) = \exp(u_i(x, y))$$

Preferably, M has an empirical value of 5. It would be appreciated that M is an adjustable parameter and can take on other values apart from the value 5.

Step 3 requires calculating the mean local variance for the illumination component of the face images obtained from T successive video frames:

Divide the illumination component of each face image equally into aBb image blocks with aBb pixels contained in each block, and use B_{ij} to represent the No.j image block of the face image from the No.i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{ij})$$

Where $\text{var}(B_{ij})$ is the variance of the pixel value of the image block B_{ij} .

The method for face in-vivo detection based on illumination component, wherein Step 4 requires conducting face in-vivo detection:

Compare the Avar value obtained from Step 3 with the predefined threshold Th, if the Avar value is greater than or equal to Th, the face image in the video is an image of a real face, otherwise a face image. The threshold Th is set according to specific image quality. A lower image resolution means a lower threshold Th.

There is described hereinafter a method for face or face in vivo detection in accordance with various embodiments of the present invention.

In an embodiment of the present invention, a method for face in vivo detection comprises the following steps:

Step 1: in order to make a facial movement video, such as a human head movement video, in actual practice a voice clip or a screen text can be used to give instructions to a user, requiring the user to shake or nod his/her head at the camera. It would be appreciated that other types of facial movement such as blinking of eyes, raising of eye brows, moving of lips and the like can also be performed and captured.

Step 2: conduct face detection for each image frame captured by the camera. As a well-known technology, face detection identifies the face in photos (or video frames) and returns the location of the face. Obtain a face region from a video frame through cropping according to face detection results and scale down the face region into a 100B100 image. Use I_i to represent the face image obtained from the No.i frame through cropping and being scaled down, where i is a natural number.

Step 3: Acquire the illumination component of each face image I_j . According to the Lambertian model, the face image I_j can be denoted as:

$$I_j(x,y)=R_j(x,y)L_j(x,y) \quad (1)$$

Where R_j is reflection component, mainly describing the surface reflectance of the image scene; L_j is the illumination component, mainly describing the illumination and shadow of the image scene. Log-transform each face image I_j to acquire:

$$\begin{aligned} f_j &= \log I_j \\ &= \log R_j + \log L_j \quad (2) \end{aligned}$$

$$\triangleq v_i + u_i$$

Where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i = \log R$, $u_i = \log L$. By this time, values of v_i and u_i are unknown and the value of u_i needs to be estimated.

Compute DCT (discrete cosine transform) for f_i , i.e.:

$$F_i(s, t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x, y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (3)$$

$$\text{Where } C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{Others} \end{cases}, \quad C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{Others} \end{cases} \quad (4)$$

Where N is the length and width of the image, and the high frequency coefficient in $F_i(s, t)$ is set at 0, i.e.:

$$F'_i(s, t) = \begin{cases} F_i(s, t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases} \quad (5)$$

Where M is a parameter to be defined, which is generally set at 5,

Compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x, y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s, t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Formula from (3) to (6) actually demonstrate the low frequency filtering for the face image f_i over the log-domain via DCT (discrete cosine transform).

It would be appreciated that in other embodiments, it can be synchronized with surface shading algorithms, ray cast and the like. It would also be appreciated that M is an adjustable parameter and can take on other values apart from the value 5.

According to a large number of existing researches, as the illumination component in images varies slowly, the low frequency component can be used to estimate the illumination component. Therefore, f'_i can be used as estimation of the illumination component, i.e.:

$$u_i(x,y) \approx f'_i(x,y) \quad (7)$$

Then the illumination component of the image domain can be obtained by computing inverse logarithmic transformation (exponential transformation), i.e.:

$$I_i(x,y) = \exp(u_i(x,y)) \quad (8)$$

Step 4: calculate the mean local variance for the illumination component of the face images obtained from T successive video frames.

Divide the illumination component of each face image equally into 10B10 image blocks with 10B10 pixels contained in each block. Use $B_{i,j}$ to denote the No.j image block of the face image from the No.i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{10 \times 10} \text{var}(B_{i,j})$$

Where $\text{var}(B_{i,j})$ is the variance of the pixel value of the image block $B_{i,j}$. In the present embodiment, T is set at 100.

Step 5: conduct face in-vivo detection.

Each face has a unique three-dimensional geometric structure (e.g. distinct unevenness can be seen around nose, cheekbones, mouth, and eyes), therefore when a person shake or nod his/her head, the regional shadow on his/her face will experience significant changes, which are properly recorded as the illumination volume L_i . A photo has a smooth surface, therefore moving the photo will not lead to significant changes in the regional shadow. As a result, the mean local variance Avar can be used to distinguish between a real face and a face image. If the Avar value is greater than the preset threshold T_h , the face image in the video can be considered as a real face, otherwise it is merely a face image. The threshold T_h is set according to specific image type and image quality. A lower image resolution means a lower threshold T_h .

In accordance with another aspect of the present invention, there is described a system

for face detection in accordance with an embodiment of the present invention. The system comprises an acquisition unit configured to capture a facial movement by video and to process the video to obtain a plurality of facial images from a plurality of successive video frames; a calculation unit configured to render each facial image obtained using the Lambertian model, compute discrete cosine transform (DTC) to obtain an illumination component of each facial image, and calculate the mean local variance for the illumination components of the facial images; and a determination unit configured to compare the mean local variance with a predetermined threshold (Th) to determine whether the facial image is an image of a real face.

Each of the facial images obtained is denoted as I_i , where i is a natural number.

The illumination component of each face image I_i , which, according to the Lambertian model, can be denoted as:

$$I_i(x,y)=R_i(x,y)L_i(x,y)$$

where R_i is the reflection component, representing the surface reflectance of the facial image; L_i is the illumination component, representing the illumination and shadow of the facial image, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_i , to obtain:

$$\begin{aligned} f_i &= \log I_i \\ &= \log R_i + \log L_i \\ &\triangleq v_i + u_i \end{aligned}$$

where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i=\log R$, $u_i=\log L$, compute DCT for f_i , i.e.:

$$F_i(s,t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x,y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

$$\text{Where } C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{others} \end{cases}, \quad C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{others} \end{cases}$$

where N is the length and width of the image, and the high frequency coefficient

of $F_i(s,t)$ is set at 0. i.e.:

$$F'_i(s,t) = \begin{cases} F_i(s,t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

where M is a parameter to be defined, which is generally set at 5, compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x,y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s,t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Take f'_i as the estimation of the illumination component, i.e.:

$$u_i(x,y) \approx f'_i(x,y) \quad (7)$$

Then the illumination component of the facial image can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x,y) = \exp(u_i(x,y))$$

It would be appreciated that in other embodiments, it can be synchronized with surface shading algorithms, ray cast and the like. It would also be appreciated that M is an adjustable parameter and can take on other values apart from the value 5.

The step of calculating the mean local variance for the illumination components of the face images obtained from T successive video frames comprises:

dividing the illumination component of each face image equally into $a \times b$ image blocks with $a \times b$ pixels contained in each block, and use $B_{i,j}$ to represent the No. j image block of the face image from the No. i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{i,j})$$

where $\text{var}(B_{i,j})$ is the variance of the pixel value of the image block $B_{i,j}$.

In the step of comparing the mean local variance ($Avar$) obtained with a predetermined threshold (Th), if the $Avar$ value is greater than or equal to Th , the face image in the video is an image of a real face. If the $Avar$ value is less than Th , the face image in the

video is not an image of a real face. The threshold (T_h) is set according to specific image quality, whereby a lower image resolution means a lower threshold T_h .

The technical effects of the present invention lie in that the detection method and system of the present invention can distinguish between a real face and a face image safely, and during detection only requires a user to perform a facial movement casually, such as to move his/her head casually, instead of making different movements as strictly required at specific times, offering a more friendly user experience. As the present invention does not rely on the detection method based on facial feature points, several deficiencies such as lower accuracy and complex calculation caused by the detection method based on facial feature points are avoided. The present invention does not involve three-dimensional face reconstruction, hence achieving higher calculation speed and performing real-time processing.

Advantageously, the present invention focuses on face in vivo detection based on the illumination information in a face image rather than relying on complex three-dimensional reconstruction and is based on facial feature points.

It is to be understood that the above embodiments have been provided only by way of exemplification of this invention, and that further modifications and improvements thereto, as would be apparent to persons skilled in the relevant art, are deemed to fall within the broad scope and ambit of the present invention described herein. It is further to be understood that features from one or more of the described embodiments may be combined to form further embodiments of the invention.

CLAIMS

1. A method for face detection, comprising:

capturing a facial movement by video and processing the video to obtain a plurality of facial images from a plurality of successive video frames;

using the Lambertian model to render each facial image obtained and computing discrete cosine transform (DTC) to obtain an illumination component of each facial image;

calculating the mean local variance for the illumination components of the facial images; and

comparing the mean local variance with a predetermined threshold (T_h) to determine whether the facial image is an image of a real face.

2. The method according to claim 1, wherein each of the facial images obtained is denoted as I_i , where i is a natural number.

3. The method according to claim 2, wherein the illumination component of each face image I_i , which, according to the Lambertian model, can be denoted as:

$$I_i(x,y)=R_i(x,y)L_i(x,y)$$

where R_i is the reflection component, representing the surface reflectance of the facial image; L_i is the illumination component, representing the illumination and shadow of the facial image, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_i , to obtain:

$$\begin{aligned}
f_i &= \log I_i \\
&= \log R_i + \log L_i \\
&\triangleq v_i + u_i
\end{aligned}$$

where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i = \log R$, $u_i = \log L$, compute DCT for f_i , i.e.:

$$F_i(s, t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x, y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

$$\text{Where } C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{Others} \end{cases}, \quad C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{Others} \end{cases}$$

where N is the length and width of the image, and the high frequency coefficient of $F_i(s, t)$ is set at 0, i.e.:

$$F'_i(s, t) = \begin{cases} F_i(s, t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

where M is a parameter to be defined, which is generally set at 5, compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x, y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s, t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Take f'_i as the estimation of the illumination component, i.e.:

$$u_i(x, y) \approx f'_i(x, y) \quad (7)$$

Then the illumination component of the facial image can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x, y) = \exp(u_i(x, y))$$

4. The method according to claim 3, wherein M has an empirical value of 5.

5. The method according to claim 4, wherein the step of calculating the mean local variance for the illumination components of the face images obtained from T successive video frames comprises:

dividing the illumination component of each face image equally into $a \times b$ image blocks with $a \times b$ pixels contained in each block, and use $B_{i,j}$ to represent the No. j image block of the face image from the No. i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{i,j})$$

where $\text{var}(B_{i,j})$ is the variance of the pixel value of the image block $B_{i,j}$.

6. The method according to claim 5, wherein in the step of comparing the mean local variance ($Avar$) obtained with a predetermined threshold (Th), if the $Avar$ value is greater than or equal to Th , the face image in the video is an image of a real face.

7. The method according to claim 6, wherein if the $Avar$ value is less than Th , the face image in the video is not an image of a real face.

8. The method according to any one of the preceding claims, wherein the threshold (Th) is set according to specific image quality, whereby a lower image resolution means a lower threshold Th .

9. A system for face detection, comprising:

an acquisition unit configured to capture a facial movement by video and to process the video to obtain a plurality of facial images from a plurality of successive video frames;

a calculation unit configured to render each facial image obtained using the Lambertian model, compute discrete cosine transform (DTC) to obtain an illumination component of each facial image, and calculate the mean local variance for the illumination components of the facial images; and

a determination unit configured to compare the mean local variance with a predetermined threshold (Th) to determine whether the facial image is an image of a real face.

10. The system according to claim 9, wherein each of the facial images obtained is

denoted as I_i , where i is a natural number.

11. The system according to claim 10, wherein the illumination component of each face image I_i , which, according to the Lambertian model, can be denoted as:

$$I_i(x,y)=R_i(x,y)L_i(x,y)$$

where R_i is the reflection component, representing the surface reflectance of the facial image; L_i is the illumination component, representing the illumination and shadow of the facial image, and (x,y) represents the coordinates of the pixels in the image; log-transform the face image I_i , to obtain:

$$\begin{aligned} f_i &= \log I_i \\ &= \log R_i + \log L_i \\ &\triangleq v_i + u_i \end{aligned}$$

where f_i , v_i and u_i respectively represent the value of I , R and L over the log-domain, i.e. $v_i=\log R$, $u_i=\log L$, compute DCT for f_i , i.e.:

$$F_i(s,t) = \frac{1}{4} C_s C_t \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f_i(x,y) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right)$$

Where $C_s = \begin{cases} 1/\sqrt{2}, & \text{if } s = 0 \\ 1, & \text{others} \end{cases}$, $C_t = \begin{cases} 1/\sqrt{2}, & \text{if } t = 0 \\ 1, & \text{others} \end{cases}$

where N is the length and width of the image, and the high frequency coefficient of $F_i(s,t)$ is set at 0, i.e.:

$$F'_i(s,t) = \begin{cases} F_i(s,t) & s < M \text{ and } t < M \quad \text{low frequencies} \\ 0 & s \geq M \text{ and } t \geq M \quad \text{high frequencies (others)} \end{cases}$$

where M is a parameter to be defined, which is generally set at 5, compute the inverse DCT (discrete cosine transform) for the adjusted frequency domain coefficient F' , i.e.:

$$f'_i(x,y) = \frac{1}{4} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} C_s C_t F(s,t) \cos\left(s\pi \frac{2y+1}{2N}\right) \cos\left(t\pi \frac{2x+1}{2N}\right) \quad (6)$$

Take f'_i as the estimation of the illumination component, i.e.:

$$u_i(x,y) \approx f'_i(x,y) \quad (7)$$

Then the illumination component of the facial image can be obtained via inverse logarithmic transformation, i.e.:

$$L_i(x, y) = \exp(u_i(x, y))$$

12. The system according to claim 11, wherein M has an empirical value of 5.

13. The system according to claim 12, wherein the step of calculating the mean local variance for the illumination components of the face images obtained from T successive video frames comprises:

dividing the illumination component of each face image equally into aBb image blocks with aBb pixels contained in each block, and use B_{ij} to represent the No.j image block of the face image from the No.i frame, therefore the mean local variance for T successive video frames is:

$$Avar = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^{a \times b} \text{var}(B_{i,j})$$

where $\text{var}(B_{i,j})$ is the variance of the pixel value of the image block $B_{i,j}$.

14. The system according to claim 13, wherein in the step of comparing the mean local variance (Avar) obtained with a predetermined threshold (Th), if the Avar value is greater than or equal to Th, the face image in the video is an image of a real face.

15. The system according to claim 14, wherein if the Avar value is less than Th, the face image in the video is not an image of a real face.

16. The system according to any one of claims 9 to 15, wherein the threshold (Th) is set according to specific image quality, whereby a lower image resolution means a lower threshold Th.