

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
10 January 2002 (10.01.2002)

PCT

(10) International Publication Number
WO 02/03709 A2

(51) International Patent Classification⁷: **H04N 7/26**

(21) International Application Number: PCT/US01/20115

(22) International Filing Date: 21 June 2001 (21.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/608,989 30 June 2000 (30.06.2000) US

(71) Applicants (*for all designated States except US*): **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052 (US). **INDIAN INSTITUTE OF TECHNOLOGY** [IN/IN]; Kharagpur 721 302 (IN).

(72) Inventors; and

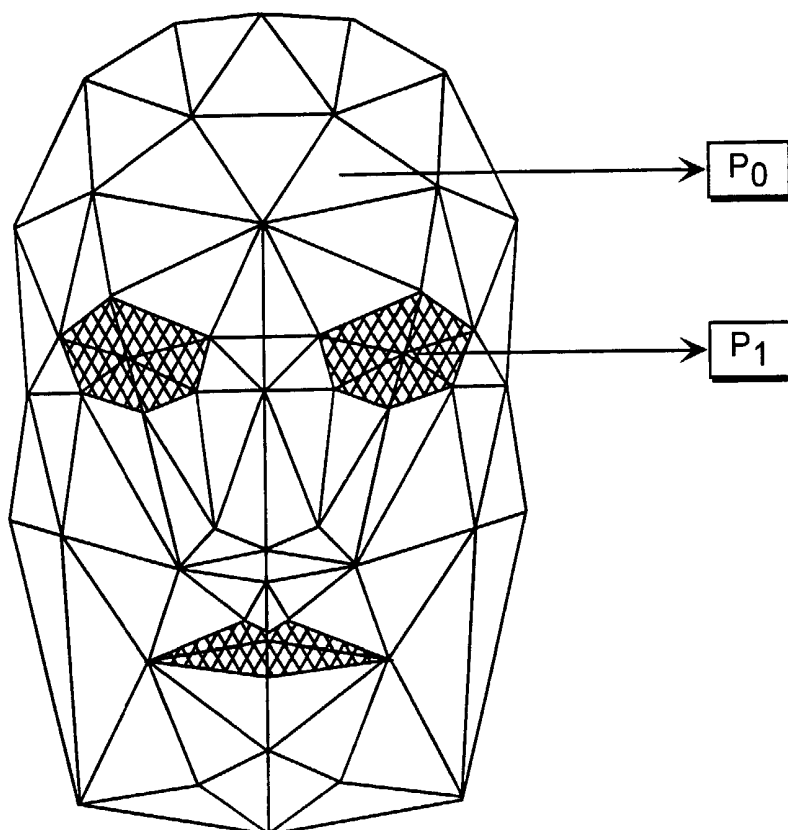
(75) Inventors/Applicants (*for US only*): **SENTUPTA, Somnath** [IN/IN]; Qr. No. 2BR-F-18, IIT Campus, Kharagpur 721 302 (IN). **SURYANARAYANA, Rama** [IN/IN]; H. No. 9-1-33/35, Prashant Nagar, Lungar House, Hyderabad 500 008 (IN).

(74) Agents: **MALLIE, Michael, J.** et al.; Blakely Sokoloff Taylor & Zafman, 12400 Wilshire Boulevard, 7th Floor, Los Angeles, CA 90025 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

[Continued on next page]

(54) Title: METHOD OF VIDEO CODING THE MOVEMENT OF A HUMAN FACE FROM A SEQUENCE OF IMAGES



(57) Abstract: Briefly, in accordance with one embodiment, a technique for coding the movement of a head or face from a sequence of images is disclosed. A variety of potential alternative embodiments are discussed.

WO 02/03709 A2



(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

**METHOD OF VIDEO CODING THE MOVEMENT OF A HUMAN FACE FROM A
SEQUENCE OF IMAGES**

BACKGROUND

The present disclosure is related to video coding and, more particularly, to coding
5 the movement of a head from a sequence of images.

As is well-known, motion estimation is a common or frequently encountered
problem in digital video processing. A number of approaches are known and have been
employed. One approach, for example, identifies the features located on the object and
10 tracks the features from frame to frame, as described for example in "Two-View Facial
Movement Estimation" by H. Li and R. Forchheimer, IEEE Transactions on Circuits and
Systems for Video Technology, Vol. 4, No. 3, pp. 276-287, June, 1994. In this approach,
the features are tracked from the two-dimensional correspondence between successive
frames. From this correspondence, the three-dimensional motion parameters are
15 estimated. Another approach estimates the motion parameters from an optical flow and
affine motion model. See, for example, "Analysis and Synthesis of Facial Image
Sequences in Model-Based Coding," by C.S. Choi, K. Aizawa, H. Harashima and T.
Takeve, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 4, No. 3,

pp. 257-275, June, 1994. This optical flow approach estimates the motion parameters without establishing a two-dimensional correspondence. This latter approach, therefore, tends to be more robust and accurate, but imposes a computational load that is heavier typically. A need, therefore, exists for an approach that is more accurate than the two-dimensional correspondence approach, but that is computationally less burdensome than the optical flow and affine motion model.

BRIEF DESCRIPTION OF THE DRAWINGS

The subject matter regarded as the invention is particularly pointed out and distinctly claimed in the concluding portion of the specification. The invention, however, both as to organization and method of operation, together with objects, features, and advantages thereof, may best be understood by reference to the following detailed description when read with the accompanying drawings in which:

FIG. 1 is a schematic diagram illustrating a three-dimensional (3D) model applied to a human face with planar triangular patches, such as may be employed in an embodiment in accordance with the present invention;

FIG. 2 is a flowchart of an embodiment in accordance with the present invention.

5

DETAILED DESCRIPTION

In the following detailed description, numerous specific details are set forth in
10 order to provide a thorough understanding of the invention. However, it will be
understood by those skilled in the art that the present invention may be practiced without
these specific details. In other instances, well-known methods, procedures, components
and circuits have not been described in detail so as not to obscure the present invention.

15 As previously described, motion estimation is a common problem in video image
processing. However, state of the art techniques such as previously described, for
example, suffer from some disadvantages. For example, the previously described
technique, referred to here as the "two-dimensional correspondence approach," although

computationally less burdensome, seems to be prone to errors due to mismatches of the two-dimensional correspondences. Another approach, referred to here as the "optical flow and affine motion model," such as described in "3-D Motion Estimation and Wireframe Adaptation Including

- 5 Photometric Effects for Model-Based Coding of Facial Image Sequences", by G.Bozdagi, A. Murat Tekalp and L. Onural, IEEE Transactions on CSVT, Vol.4, No.3, pp.246-256, June 1994, although more accurate and robust, is typically computationally burdensome. Therefore, a need exists for an approach that is more accurate than the former, but less computationally burdensome than the latter.

10

- In this particular context, the motion that is being tracked or coded is the movement of a head or face in a sequence of images. Having the ability to track this motion and coding it may be desirable for a number of reasons. As just a few examples, this may be desirable in video conferencing, where a camera at one end may transmits
- 15 the appropriate motion or movement of face to a display at the other end. However, the communications channel by which this video conferencing may take place sometimes has a relatively low or limited bandwidth, so that only a limited amount of signal information may be communicated in real-time.

An embodiment of a method of video coding a movement of human head or face from a sequence of images includes the following. A limited number of feature points are selected from an image of the face whose movement is to be video coded. Using at least
5 two images or frames from the sequence, changes in the intensity of selected feature points, such as spatio-temporal rates of change, are estimated. Using the feature points and the estimated rates, translation and rotation parameters of the face are then estimated. The estimated translation and rotation parameters are coded and/or transmitted across the communications channel. It is noted, of course that instead of
10 communicating the coded signal information, it may, alternatively, be stored and read from memory for later use, or used in some other way other than by transmitting it.

Although the invention is not limited in scope in this respect, in this particular embodiment, the face is coded from at least one of the images or frames by employing a
15 three-dimensional (3D) based coding technique to produce what shall be referred to here as a 3D model. Movement of the face from at least two, typically sequential, images of the sequence is estimated using this 3D model of the face or head. In particular, as

shall be described in more detail hereinafter, the movement of the face is estimated by treating the 3D model of the head as a rigid body in the sequence of images.

In this embodiment, although the invention is not limited in scope in this respect, the 3D model applied comprises planar triangular patches. This illustrated, for example, in FIG. 1. As FIG. 1 illustrates, these triangular patches, in this particular embodiment in accordance with the invention, are divided into two classes, one class in which local motion is more significant, such as, for example, the triangular patches covering eyes, eyebrows, or mouth, denoted here p_l , and one class in which global motion is more significant, denoted here by the p_g . FIG. 1 illustrates the two classes of triangles, the shaded of triangles belonging to p_l and unshaded triangles belonging to p_g .

In this embodiment, a limited number of feature points are selected from an image of the head. In this embodiment, enough feature points are selected from different triangular patches to obtain the desired amount of accuracy or robustness without being computationally burdensome. Furthermore, a weighting factor is assigned to each feature point, depending upon the class of triangular patch to which it belongs. The

weighting factor assigned to a feature point selected from the i^{th} triangular patch is given by the following relationship.

$$W_{pi} = \begin{cases} W_g, & \text{for all } i \in \rho_g \\ W_l, & \text{for all } i \in \rho_l \end{cases}$$

where W_g is greater than W_l .

The weighting factors are used in the Least Mean Square estimation of the global motion parameters in this particular embodiment, as described in more detail later, and there, the facial regions contributing more to the global motion have more weighting factors than the ones predominantly contributing to local motion; however, the invention is not restricted in scope to this embodiment. For example, other estimation approaches other than Least Mean Square may be employed and other approaches to employing weighting may be employed, or, alternatively, weighting may not necessarily be employed in alternative embodiments. For this embodiment, the range of the weighting factors were determined from experimentation, although, again, the invention is not restricted in scope to this particular range of weights. Here, nonetheless, W_g varies in the range of approximately 0.6 to approximately 0.9 and W_l varies in the range of approximately 0.3 to approximately 0.1.

Once feature points are selected, the rate of change of intensity of the selected feature points is estimated from the sequence of images. It is noted that it takes at least two images to estimate a rate of change; however, in this embodiment a rate of change is calculated for each pair of immediately sequential images in the sequence. It is also noted that a distinguishing feature of this approach is the selection of a limited number of feature points, thereby reducing the computational burden of this approach.

The relationship between rate of change in intensity at the selected feature points and estimating the translation and rotation of the face is as follows. The gradient between two consecutive or immediately sequential frames is described as follows.

$$I_{xk} V_{xk} + I_{yk} V_{yk} + I_{tk} = 0 \dots\dots\dots(1)$$

where I_{xk} , I_{yk} , and I_{tk} are the rates of change at a selected pixel between the two frames k and $(k+1)$ in the x -, y - and the temporal directions respectively and V_{xk} , V_{yk} are optical flow fields in the x and y directions, respectively. The I_{xk} and I_{yk} are determined by the intensity gradients of the neighboring pixels in the same frame and I_{tk} is measured from the intensity gradient at substantially the same spatial position between

consecutive frames. The equation is based on an assumption of
brightness constancy for moving objects in the successive frames.

Likewise, the formula for small motion of a rigid body is given by the following equation.

$$V = R P + T \dots\dots\dots(2)$$

where P is a three-dimensional position vector, vector V represents the
velocity of a point on the rigid body, matrix R represents the angular
velocity, and vector T represents the translation of the rigid body . R, the
angular velocity, is given by the following 3-by-3 matrix

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

where $r_{11} = c_{\alpha}c_{\beta} - 1$, $r_{12} = c_{\alpha}s_{\beta}s_{\gamma} - s_{\alpha}c_{\gamma}$, $r_{13} = c_{\alpha}s_{\beta}c_{\gamma} - c_{\alpha}s_{\gamma}$,

$r_{21} = s_{\alpha}s_{\beta}$, $r_{22} = s_{\alpha}s_{\beta}s_{\gamma} - c_{\alpha}c_{\gamma} - 1$, $r_{23} = s_{\alpha}s_{\beta}c_{\gamma} - c_{\alpha}s_{\gamma}$, $r_{31} = -s_{\beta}$, $r_{32} = c_{\beta}s_{\gamma}$,

$r_{33} = c_{\beta}c_{\gamma} - 1$ and α , β , and γ are rotations about x, y, and z axes, respectively, and

c and s denote the cosine and the sine of the angles.

Under the assumption of orthographic projection of the human face, for this particular embodiment, V_{xk} and V_{yk} are considered to be the optical flow fields with the z-directional component assumed to be zero. The following linearized estimation equation may, therefore, be derived from equation (2) above by equating the x- and the y-
 5 directional components of the velocities and then using these relations in equation (1) to evaluate I_{TK} as

$$H_K = F_K A$$

where H_K is $-I_{TK}$, A is $[r_{11} \ r_{12} \ r_{13} \ r_{21} \ r_{22} \ r_{23} \ T_x \ T_y]$, and F_K is

$$10 \quad [x_k I_{xk} \ y_k I_{xk} \ z_k I_{xk} \ x_k I_{yk} \ y_k I_{yk} \ z_k I_{yk} \ I_{xk} \ I_{yk}]$$

The global estimation parameters, here, the translation and rotation signal information, may be obtained by solving the previous equation using a Least Mean Square approach, although, of course, the invention is not limited in scope in this respect, and other approaches, such as least absolute value, may be employed. This is done for each
 15 feature point and then the values for a rotation and translation that give the least overall mean square error are employed, again, for this particular embodiment.

FIG. 2 is a flowchart illustrating an embodiment 200 of a method of video coding the movement of a human face from a sequence of images in accordance with the invention, although, of course, this is just one example of an embodiment, and many others are possible within the scope of the present invention. At block 210, the face or head from a particular or selected image is approximated using triangular patches, such as illustrated, for example, in FIG. 1. At block, 220, a limited number of feature points on the face are selected, in this particular embodiment, one feature point from each triangular patch, although, of course, the invention is not limited in scope in this respect. At block 230, for each pair of successive images or frames in the sequence of images, spatio-temporal rates of change in intensity at the selected feature points are estimated. At block 240, estimates of translation and rotation for the face are made using the feature points and the estimated spatio-temporal rates of change for each pair of successive images in this particular embodiment. At block 250, these estimates are then coded and transmitted across a transmission medium or communications channel so that at the far end, the estimates may be employed to reproduce or represent movement of a representation of the face.

It will, of course, be understood that, although particular embodiments have just been described, the invention is not limited in scope to a particular embodiment or implementation. For example, one embodiment may be in hardware, whereas another embodiment may be in software. Likewise, an embodiment may be in firmware, or any
5 combination of hardware, software, or firmware, for example. Likewise, although the invention is not limited in scope in this respect, one embodiment may comprise an article, such as a storage medium. Such a storage medium, such as, for example, a CD-ROM, or a disk, may have stored thereon instructions, which when executed by a system, such as a host computer or computing system or platform, or an imaging system, may result in
10 a method of video coding the movement of a human face from a sequence of images in accordance with the invention, such as, for example, one of the embodiments previously described. Likewise, a hardware embodiment may comprise an imaging system including an imager and a computing platform, such as one adapted to perform or execute coding in accordance with the invention, for example.

15

While certain features of the invention have been illustrated and detailed herein, many modifications, substitutions, changes and equivalents will now occur to those

skilled in the art. It is, therefore, to be understood that the appended claims are intended to cover all such modifications and changes as fall within the true spirit of the invention.

Claims:

1. A method of video coding the movement of a face from a sequence of images
comprising:

selecting a limited number of feature points from an image of the face whose
5 movement is to be video coded;

estimating spatio-temporal rates of change in intensity at the selected feature
points using at least two images from the sequence;

estimating the translation and rotation of the face using the feature points and
using the estimated spatio-temporal rates; and
10 coding the estimated translation and rotation.
2. The method of claim 1, wherein prior to selecting the feature points, for a
particular image, the face is approximated using triangular planar patches.
- 15 3. The method of claim 2, wherein a feature point is selected from each planar
triangular patch.

4. The method of claim 1, wherein the translation and rotation of the face are estimated from the feature points and from the estimated spatio-temporal rates using a least mean square estimation method.

5 5. A method of video coding the movement of a head from a sequence of images comprising:

coding the head from at least one of the images based, at least in part, on a limited number of selected feature points employing a three-dimensional based coding technique to produce a three-dimensional (3D) model; and

10 estimating the movement of the head in other images of the sequence using the 3D model of the head.

6. The method of claim 5, wherein the movement of the head is estimated by treating the 3D model of the head as a rigid body in the sequence of images.

15

7. The method of claim 6, wherein the movement of the head is estimated as translations and rotations based at least in part on estimates of spatio-temporal rates of change in intensity at the selected feature points.

8. A system comprising:
- an imager; and a computing platform;
- said imager and computing platform being coupled to communicate electronically;
- 5 wherein said computing platform being adapted so that, in operation, the movement of a face from a sequence of images is coded by:
- selecting a limited number of feature points from an image of the face whose movement is to be video coded;
- estimating spatio-temporal rates of change in intensity at the selected feature
- 10 points using at least two images from the sequence;
- estimating the translation and rotation of the face using the feature points and using the estimated spatio-temporal rates; and
- coding the estimated translation and rotation.
- 15 9. The system of claim 8, wherein said computing platform is adapted so that, in operation, the movement of a face from a sequence of images is further coded by, prior to selecting the feature points, for a particular image, approximating the face using triangular planar patches.

10. The system of claim 9, wherein said computing platform is adapted so that, in operation, the movement of a face from a sequence of images is further coded by selecting a feature point from each planar triangular patch.

5

11. The system of claim 8, wherein said computing platform is adapted so that, in operation, the movement of a face from a sequence of images is further coded by estimating the translation and rotation of the face from the feature points and from the estimated spatio-temporal rates using a least mean square estimation method.

10

12. A system comprising:
an imager; and a computing platform;
said imager and computing platform being coupled to communicate electronically;
wherein said computing platform being adapted so that, in operation, the movement of a

15 head from a sequence of images is represented by:

coding the head from at least one of the images based, at least in part, on a limited number of selected feature points employing a three-dimensional based coding technique to produce a three-dimensional (3D) model; and

estimating the movement of the head in other images of the sequence using the
3D model of the head.

13. The system of claim 12, wherein said computing platform is adapted so that, in
5 operation, the movement of the head in other images of the sequence is estimated by
treating the 3D model of the head as a rigid body in the sequence of images.

14. The system of claim 13, wherein said computing platform is adapted so that, in
operation, the movement of the head is estimated as translations and rotations based at
10 least in part on estimates of spatio-temporal rates of change in intensity at the selected
feature points.

15. An article comprising: a storage medium, said storage medium having stored
thereon instructions, said instructions, when executed by a computing platform, resulting
15 in the movement of a face from a sequence of images being coded by:

selecting a limited number of feature points from an image of the face whose
movement is to be video coded;

estimating spatio-temporal rates of change in intensity at the selected feature

points using at least two images from the sequence;

estimating the translation and rotation of the face using the feature points and

using the estimated spatio-temporal rates; and

5 coding the estimated translation and rotation.

16. The article of claim 15, wherein said instructions, when executed, further result in,
prior to selecting the feature points, for a particular image, approximating the face using
triangular planar patches.

10

17. The article of claim 16, wherein said instructions, when executed, further result in
selecting a feature point from each planar triangular patch.

18. The article of claim 15, wherein said instructions, when executed, further result in,
15 estimating the translation and rotation of the face from the feature points and from the
estimated spatio-temporal rates using a least mean square estimation method.

19. An article comprising: a storage medium, said storage medium having stored thereon instructions, said instructions, when executed by a computing platform, result in the movement of a head from a sequence of images being represented by:

coding the head from at least one of the images based, at least in part, on a limited number of selected feature points employing a three-dimensional based coding technique to produce a three-dimensional (3D) model; and estimating the movement of the head in other images of the sequence using the 3D model of the head.

20. The article of claim 19, wherein said instructions, when executed, further result in the movement of the head in other images of the sequence being estimated by treating the 3D model of the head as a rigid body in the sequence of images.

21. The article of claim 20, wherein said instructions, when executed, result in the movement of the head being estimated as translations and rotations based at least in part on estimates of spatio-temporal rates of change in intensity at the selected feature points.

1/2

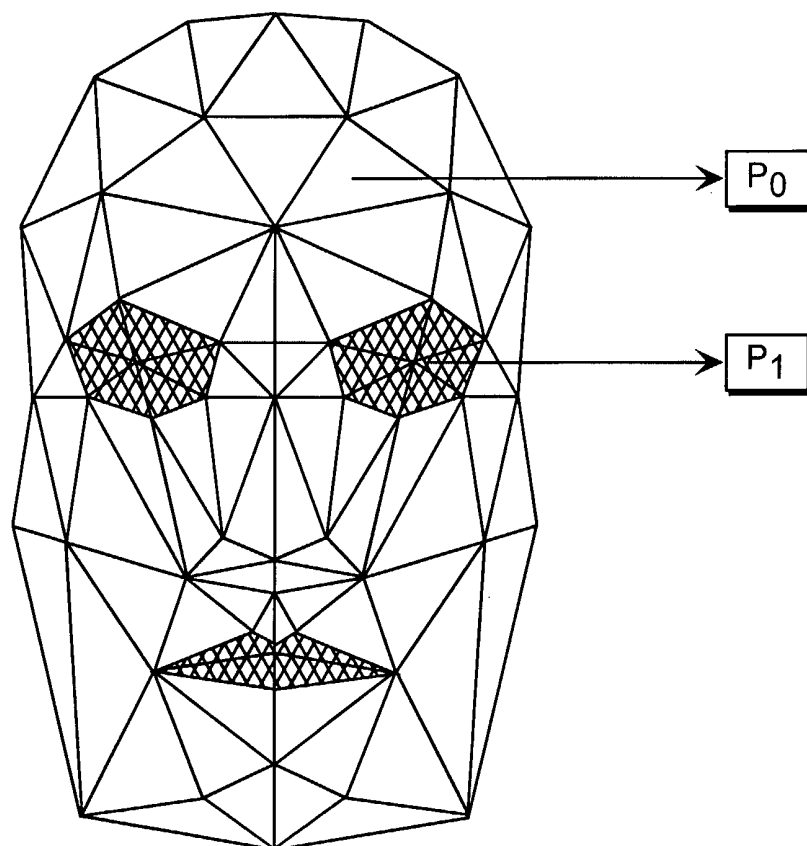


FIG. 1

2/2

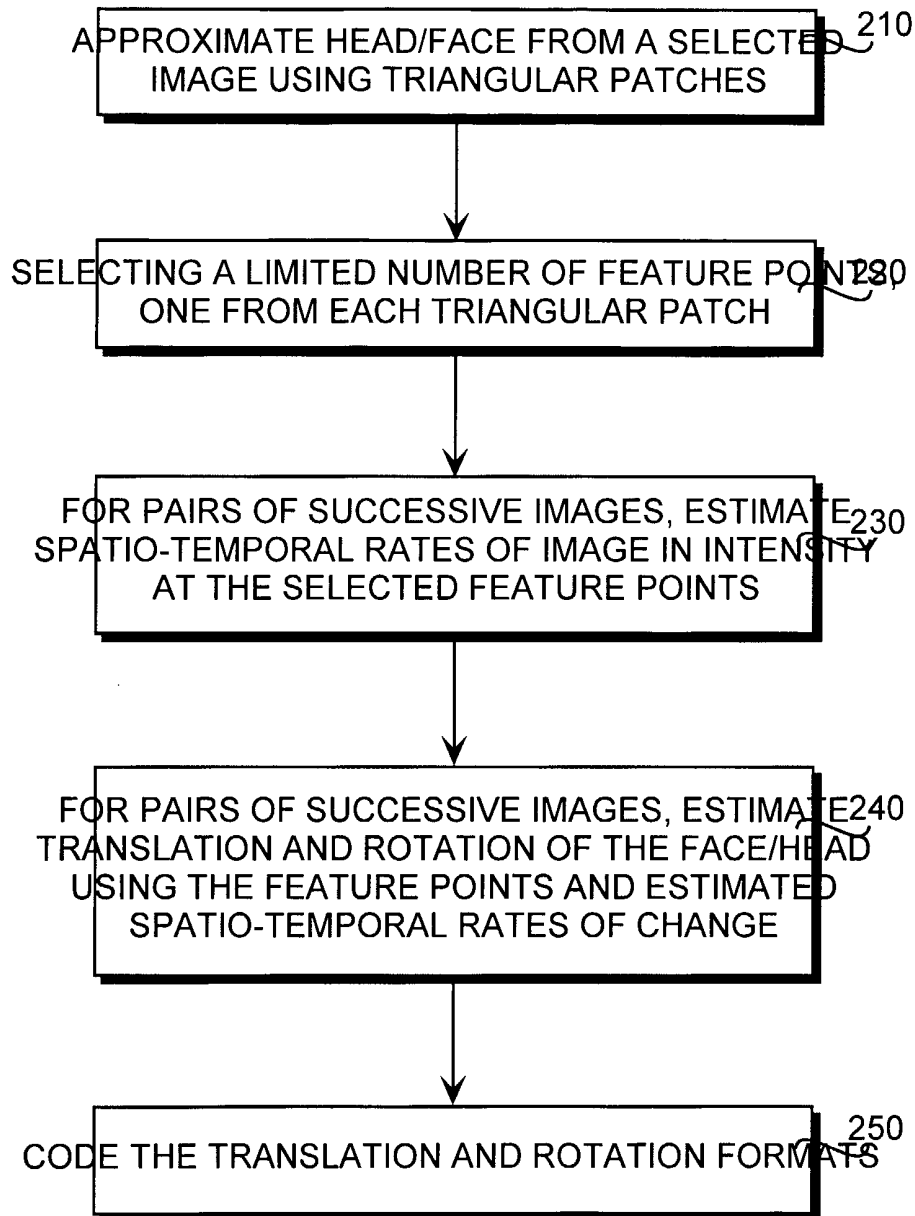
200
↙

FIG. 2