



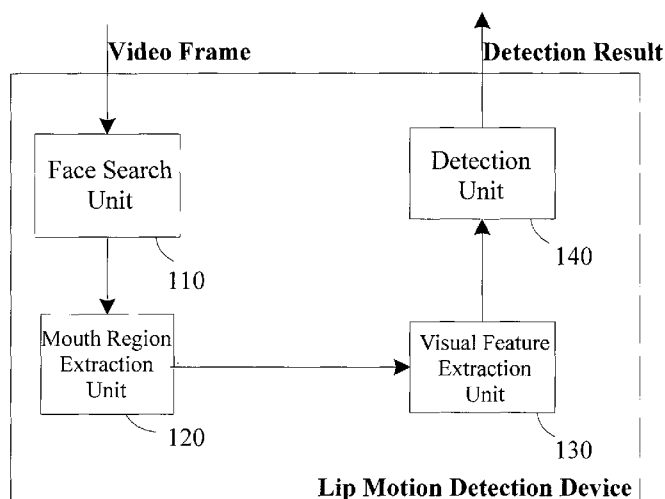
- (51) **International Patent Classification:**
G06T 7/20 (2006.01) *G06T 7/00* (2006.01)
- (21) **International Application Number:**
PCT/JP2012/057677
- (22) **International Filing Date:**
19 March 2012 (19.03.2012)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
201110077483.1 18 March 2011 (18.03.2011) CN
- (71) **Applicant (for all designated States except US):** SHARP KABUSHIKI KAISHA [JP/JP]; 22-22, Nagaike-cho, Abeno-ku, Osaka-shi, Osaka, 5458522 (JP).
- (72) **Inventor; and**
- (75) **Inventor/Applicant (for US only):** YAN, Wang.
- (74) **Agent:** HARAKENZO WORLD PATENT & TRADE-MARK; Daiwa Minamimorimachi Building, 2-6, Tenjin-bashi 2-chome Kita, Kita-ku, Osaka-shi, Osaka, 5300041 (JP).

(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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) **Title:** DEVICE AND METHOD FOR LIP MOTION DETECTION**FIG. 1**

(57) **Abstract:** A device for video-based lip motion detection is provided, which comprises: a face search unit adapted for searching a face in an input video frame; a mouth region extraction unit adapted for extracting a mouth region from the searched face; a visual feature extraction unit adapted for extracting at least one of a gradient on spatial-temporal plane and a Local Binary Pattern (LBP) code on spatial-temporal plane for each pixel in a spatial-temporal window and extracting a visual feature of the mouth region based on the extraction result; and a detection unit adapted for detecting lip motion based on the extracted visual feature of the mouth region. A method for video-based lip motion detection is also provided. According to the present invention, the detection accuracy can be improved without re-training or adaptation for a particular user.

- 1 -

DESCRIPTION

TITLE OF INVENTION:

DEVICE AND METHOD FOR LIP MOTION DETECTION

TECHNICAL FIELD

5 The invention relates to video processing, and more particularly, to a device and method for video-based lip motion detection.

BACKGROUND ART

10 In a noisy environment, it is difficult to accurately detect a speech segment based on audio signal only. The speech segment may sometimes be improperly cropped and/or appended with noise. As a consequence, the speech recognition accuracy may be degraded.

15 It is known that lip motion is a good indicator of speech. US7343289B2 discloses a system and method for audio/video speaker detection. The method is directed to detect the speaker, i.e. the subject of lip motion, based on both visual and audio information. In particular, the disclosed method includes the
20 following steps of: finding face from video frame; finding and extracting the mouth region; extracting the mouth openness with Linear Discriminant Analysis (LDA) as visual feature; extracting the energy of the audio signal corresponding to the

- 2 -

video frame as audio feature; and inputting both features into a trained Time-Delayed Neural Network (TDNN) and detecting lip motion according to the output of the TDNN.

However, the above method extracts visual feature from each frame separately. The feature includes rich information on the subject's identity and is thus somewhat individual-dependent. As a consequence, if the method is used to detect lip motion of a subject not included in the training set of the TDNN, the detection rate will be degraded significantly.

10

SUMMARY OF INVENTION

In order to solve the above technical problem, according to an aspect of the present invention, a device for video-based lip motion detection is provided, which comprises: a face search unit adapted for searching a face in an input video frame; a mouth region extraction unit adapted for extracting a mouth region from the searched face; a visual feature extraction unit adapted for extracting a visual feature of the mouth region; and a detection unit adapted for detecting lip motion based on the extracted visual feature of the mouth region.

Preferably, the visual feature extraction unit is adapted for extracting at least one of a gradient on spatial-temporal plane and a Local Binary Pattern (LBP) code on spatial-temporal plane for each pixel in a spatial-temporal window and extracting a visual feature of the mouth region based on the

25

extraction result.

Preferably, the detection unit is pre-trained with the extracted visual feature of the mouth region.

Preferably, the device for video-based lip motion detection
5 further comprises a smoothing unit adapted for smoothing the detection result of the detection unit.

Preferably, the device for video-based lip motion detection further comprises: an audio feature extraction unit adapted for extracting an audio feature corresponding to the input video
10 frame. The detection unit is adapted for detecting lip motion based on the visual feature extracted by the visual feature extraction unit and the audio feature extracted by the audio feature extraction unit.

Preferably, the detection unit is pre-trained with the
15 extracted visual feature and audio feature.

Preferably, the visual feature comprises a Local Binary Pattern on Three Orthogonal Planes (LBP-TOP)-based visual feature.

Preferably, the mouth region is a rectangle, the center of
20 the rectangle being located at the middle of the line connecting two mouth corners and the longer edges of the rectangle being parallel with the line connecting two mouth corners.

Preferably, the detection unit comprises a Support Vector Machine (SVM).

25 Preferably, the smoothing unit comprises a median filter.

Preferably, the face search unit comprises a Viola-Jones face detector.

Preferably, the mouth region extraction unit is adapted for extracting the mouth region from the searched face by using an
5 Active Shape Model (ASM).

Preferably, the visual feature extraction unit is further adapted for extracting at least one of a gradient on image plane and a Local Binary Pattern (LBP) code on image plane for each pixel in a spatial-temporal window and extracting a visual
10 feature of the mouth region based on the overall extraction result

According to another aspect of the present invention, a method for video-based lip motion detection is provided, which comprises the following steps of: searching a face in an input
15 video frame; extracting a visual feature of the mouth region; and detecting lip motion based on the extracted visual feature of the mouth region.

Preferably, the method for video-based lip motion detection further comprises: extracting at least one of a
20 gradient on spatial-temporal plane and a Local Binary Pattern (LBP) code on spatial-temporal plane for each pixel in a spatial-temporal window, prior to extracting a visual feature of the mouth region. The visual feature of the mouth region is extracted based on the extraction result.

25 Preferably, the method for video-based lip motion

detection further comprises: pre-training with the extracted visual feature of the mouth region, prior to detecting lip motion.

Preferably, the method for video-based lip motion detection further comprises: smoothing the detection result.

5 Preferably, the method for video-based lip motion detection further comprises: extracting an audio feature corresponding to the input video frame. The lip motion is detected based on the extracted visual feature and audio feature.

10 Preferably, the method further comprises: pre-training with the extracted visual feature and audio feature, prior to detecting lip motion.

15 Preferably, the visual feature comprises a Local Binary Pattern on Three Orthogonal Planes (LBP-TOP)-based visual feature.

Preferably, the mouth region is a rectangle, the center of the rectangle being located at the middle of the line connecting two mouth corners and the longer edges of the rectangle being parallel with the line connecting two mouth corners.

20 Preferably, the lip mouth is detected by using a Support Vector Machine (SVM).

Preferably, the detection result is smoothed by using a median filter.

25 Preferably, the face is searched from the input video frame by using a Viola-Jones face detector.

Preferably, the mouth region is extracted from the searched face by using an Active Shape Model (ASM).

Preferably, the method for video-based lip motion detection further comprises: extracting at least one of a
5 gradient on image plane and a Local Binary Pattern (LBP) code on image plane for each pixel in a spatial-temporal window and extracting a visual feature of the mouth region based on the overall extraction result.

According to another aspect of the present invention, a
10 speech recognition system is provided, which comprises: a microphone adapted for capturing an audio signal; a camera adapted for capturing a video signal; a device for video-based lip motion detection, adapted for detecting lip motion based on the video signal captured by the camera to obtain start time and
15 end time of the lip motion; a speech segment detector adapted for extracting a speech segment based on the audio signal captured by the microphone and the start time and the end time of the lip motion; a feature extractor adapted for extracting an audio feature from the extracted speech segment; and a speech
20 recognizer adapted for recognizing speech based on the extracted audio feature.

According to another aspect of the present invention, a video conference system is provided, which comprises: a microphone adapted for capturing an audio signal; a camera
25 adapted for capturing a video signal; a device for video-based

- 7 -

lip motion detection, adapted for detecting lip motion based on the video signal captured by the camera to obtain start time and end time of the lip motion; and a transmitter. The device for video-based lip motion detection is adapted for controlling the transmitter to transmit the audio signal captured by the microphone and the video signal captured by the camera at the start time of the lip motion and for controlling the transmitter to transmit only the video signal captured by the camera at the end time of the lip motion.

10 Preferably, the video conference system further comprises: a video frame cropper adapted for cropping video from the video signal captured by the camera. The device for video-based lip motion detection is adapted for enabling the video frame cropper and controlling the transmitter to transmit the video signal captured by the microphone and the video cropped by the video frame cropper at the start time of the lip motion and for disabling the video frame cropper and controlling the transmitter to transmit only the video signal captured by the camera at the end time of the lip motion.

20 Preferably, the video frame cropper is adapted for cropping a close-up of a speaker who is currently speaking by means of zooming.

With the present invention, it is possible to implement subject-independent lip motion detection for a training set containing a limited number of subjects. Compared with prior

- 8 -

art, the present invention enables a higher detection rate for a subject not contained in the training set. According to the present invention, the retraining or adaptation for different users for improving the detection rate is not necessary any
5 more, and thus the usability is improved.

BRIEF DESCRIPTION OF DRAWINGS

The above and other features of the present invention will be more apparent from the following detailed description with
10 reference to the figures, in which:

Fig. 1 shows a block diagram of a device for video-based lip motion detection according to an embodiment of the present invention;

Fig. 2 shows an example for calculating LBP code
15 according to an embodiment of the present invention;

Fig. 3 shows an example for extracting LBP-TOP-based feature according to an embodiment of the present invention;

Fig. 4 shows a block diagram of a device for video-based lip motion detection according to another embodiment of the
20 present invention;

Fig. 5 shows a block diagram of a device for video-based lip motion detection according to another embodiment of the present invention;

Fig. 6 illustrates a flowchart of a method for video-based
25 lip motion detection according to an embodiment of the present

invention;

Fig. 7 shows a block diagram of a video-aided speech recognition system having a device for lip motion detection according to an embodiment of the present invention;

5 (a) to (c) of Fig. 8 show signals in the speech recognition system of Fig. 7; and

Fig. 9 shows a block diagram of a video conference system having a device for lip motion detection according to an embodiment of the present invention.

10

DESCRIPTION OF EMBODIMENTS

The embodiments of the present invention will be detailed with reference to the drawings, from which the principles and implementations of the present invention will become more
15 apparent. It is to be noted that the present invention is not limited to the particular embodiments described in the following. In addition, details of well known techniques unnecessary to the present invention are omitted herein for simplicity.

20 Fig. 1 shows a block diagram of a device 10 for video-based lip motion detection according to an embodiment of the present invention. As shown in Fig. 1, the lip motion detection device 10 includes: a face search unit 110 adapted for searching a face in an input video frame; a mouth region extraction unit 120
25 adapted for extracting a mouth region from the searched face; a

- 10 -

visual feature extraction unit 130 adapted for extracting a visual feature of the mouth region; and a detection unit 140 adapted for detecting lip motion based on the extracted visual feature of the mouth region. In the following, the operations of the respective components included in the lip motion detection device 10 will be detailed.

The face search unit 110 searches a face in each input video frame. If any face is found, its position will be transferred to the mouth region extraction unit 120 as input information. On the other hand, no further process will be taken on a video frame in which no face is found. The face search unit 100 can be implemented using various known techniques for face detection and tracking, such as but not limited to: Viola-Jones face detector, Rowley face detector, meanshift tracker and particle filtering tracker.

The mouth region extraction unit 120 searches a mouth region from the face found by the face search unit 110 and extracts it from the face. In particular, for each found face, two mouth corners are searched first. Then, the mouth region is determined based on the two found mouth corners. The two mouth corners can be located by the well-known Active Shape Model (ASM). Alternatively, the Active Appearance Model (AAM) and Snakes (also known as Active Contour Model) can also be used to locate the two mouth corners. After the two mouth corners are located, a rectangular region can be determined,

- 11 -

with the center of the rectangular region located at the middle of the line connecting two mouth corners and the longer edges of the rectangular region being parallel with the line connecting two mouth corners. Then, the rectangular region is determined
5 as the mouth region. The aspect ratio of the rectangular region is preferably 3:2. However, other aspect ratios are also applicable.

Alternatively, the mouth region can be of any other shape, such as ellipse, which contains the entire outer contour of the
10 lip. In addition, these shapes are not necessarily symmetric and their centers are not necessarily coincident with the middle of the line connecting the mouth corners. Further, any shape having a large intersection with the outer contour of the lip can also be used as the mouth region.

15 The visual feature extraction unit 130 extracts a visual feature from a spatial-temporal window which contains one or more consecutive mouth regions. In an embodiment of the present invention, the visual feature is described as a Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) feature,
20 which is a spatial-temporal extension of the well-known Local Binary Pattern (LBP) feature.

Specifically, for a given pixel (x_c, y_c) , values of its P uniformly spaced neighbors on a circle which is centered at (x_c, y_c) and has a radius of R are extracted by interpolation. The
25 LBP code of the pixel is given by

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p,$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

g_c is the value of the pixel (x_c, y_c) and g_p is the value of the
 5 p -th neighbor. Fig. 2 shows an example for calculating the LBP
 code, where $P=8$ and $g_c=70$. The values of the central pixel and
 its neighbors uniformly distributed on a surrounding unit circle
 are shown in (a) of Fig. 2. First, the value of the central pixel is
 compared with the value of each of its neighbors. A neighbor
 10 can be represented as 1 if its value is not smaller than the value
 of the central pixel; otherwise it can be represented as 0. (b) of
 Fig. 2 shows the comparison result. Then, starting from the
 horizontal neighbor on the left, the comparison can be arranged,
 in a counter-clockwise order, into a binary value which is the
 15 LBP code of the central pixel. Of course, any other order or
 starting pixel is also applicable.

Besides the basic form, there are many known variations
 of the LBP code, e.g., uniform LBP, rotation invariant LBP and a
 combination thereof. Preferably, the basic LBP code with $P=8$
 20 and $R=1$ is employed. However, other LBP code variations and
 other values of P and R are also applicable.

The LBP-TOP feature is extracted from a spatial-temporal
 window which contains one or more consecutive mouth regions.

- 13 -

Fig. 3 shows an example of the extraction process. As illustrated in Fig. 3, a window (as shown in (a) of Fig. 3) is first divided into one or more spatial-temporal blocks. The LBP codes for each pixel in each block (as shown in (b) of Fig. 3) are
5 extracted with respect to its neighbors on XY, XT and/or YT planes, respectively. For each block, as shown in (c) of Fig. 3, a LBP code histogram for each plane is extracted and the histograms from one or more specific planes are then concatenated into the LBP-TOP feature of the block. The
10 histogram from the XY plane contains more information on the subject's identity, while the histograms from the XT and YT planes contain more information on motion, which is less individual-dependent. Finally, the LBP-TOP features from all blocks are concatenated into the LBP-TOP feature of the
15 spatial-temporal window.

As used in this embodiment, the spatial-temporal window comprises five consecutive mouth regions and is divided uniformly into $6 \times 4 \times 1$ (corresponding to X, Y and T axes, respectively) blocks with 50% of overlap. Only LBP histograms
20 from the XT and YT planes are used. By combining the less individual-dependent information from the XT and YT planes, the present invention works well on subjects not contained in the training set. However, other numbers of mouth regions, other types of window division and other combinations of LBP
25 histograms from different planes are also applicable.

It can be appreciated by those skilled in the art that the feature based on LBP code can be extracted by post-processing the histograms or otherwise. For example, after the histogram of each plane for each spatial-temporal block is extracted, the
5 histograms can be normalized. The normalization can be carried out separately for each plane of each block. Alternatively, the histograms of the same plane for spatially and temporally adjacent blocks, the histograms of different planes for the same block, or the histograms of different planes
10 for spatially and temporally adjacent blocks can be normalized collectively. The criterion of normalization may cause the sum, or squared sum, of the normalized vector elements to be 1. After a first normalization, a value in a histogram exceeding a certain threshold can be changed to that threshold and a further
15 normalization can be performed. Finally, the normalized histograms can be concatenated into the feature of the spatial-temporal window. In addition, after the LBP code for each pixel is calculated, instead of extracting the histogram, the LBP code(s) on one or more particular planes for all pixels of
20 each block can be considered as a feature vector which is dimension-reduced by means of sub-space analysis, such as Principal Component Analysis or Linear Discriminant Analysis, and then used as the feature of the block or plane. Finally, the LBP-TOP features of all blocks and/or all planes are
25 concatenated into the feature of the spatial-temporal window.

Also, it is understood by those skilled in the art that other features can be used to describe a visual feature, such as but not limited to spatial-temporal extension of gradient type feature. Traditional gradient type feature calculates a gradient
5 on the XY image plane and extracts a feature based on the gradient, while the spatial-temporal extension of gradient type feature calculates features on the XY, XY and/or YT planes, respectively. Then, the gradient direction histograms for the respective planes in each spatial-temporal block can be
10 extracted and normalized according to any one of normalization approaches in LBP-TOP for extracting feature. Alternatively, the gradient(s) on one or more planes for all pixels of each block can be considered as a feature vector which is dimension-reduced by means of sub-space analysis, such as
15 Principal Component Analysis or Linear Discriminant Analysis, and then used as the feature of the block or plane. Finally, the gradient features from all blocks and/or all planes are concatenated into the feature of the spatial-temporal window, thereby obtaining the visual feature of the mouth region.

20 The detection unit 140 detects lip motion based on the extracted visual feature of the mouth region. For example, the detection unit 140 can be a classifier capable of distinguishing two classes of mouth regions, i.e., mouth region with lip motion and mouth region without lip motion. There are a number of
25 known classifiers. In a preferred embodiment, a Support Vector

- 16 -

Machine (SVM) is used. However, other classifiers such as k-Nearest Neighbor classifier, Adaboost classifier, Neural Network classifier, Gaussian Process classifier and threshold classifier based on feature similarity are also applicable.

5 Preferably, before detecting the lip motion, the detection unit 140 can be pre-trained with the extracted visual feature of the mouth region. In an embodiment, such training can be achieved by assigning a label to each extracted visual feature. For example, if there is a lip motion in the mouth region
10 corresponding to a visual feature, a label of +1 is assigned to the feature, otherwise a label of -1 is assigned. Then, the detection unit 140 can be trained by using a number of known training approaches.

Fig. 4 shows a block diagram of a device 40 for video-based
15 lip motion detection according to another embodiment of the present invention. As shown in Fig. 4, the lip motion detection device 40 includes: a face search unit 410, a mouth region extraction unit 420, a visual feature extraction unit 430, a detection unit 440 and an audio feature extraction unit 450.
20 Since the face search unit 410, mouth region extraction unit 420, visual feature extraction unit 430 and detection unit 440 as shown in Fig. 4 are similar to the face search unit 110, mouth region extraction unit 120, visual feature extraction unit 130 and detection unit 140 as shown in Fig. 1, only the audio
25 feature extraction unit 450 will be detailed in the following for

simplicity.

In this embodiment, the audio feature extraction unit 450 extracts an audio feature corresponding to the input video frame. The extracted audio feature is provided to the detection unit 440 along with the visual feature extracted by the visual feature extraction unit 430. In particular, if a subject in the video is speaking and a synchronized audio is available, the audio feature extraction unit 450 can use any known audio-based speech endpoint detection methods to detect the speech part and non-speech part. If a frame falls in the speech part, there is a high probability of presence of lip motion in the mouth region in the frame. The detection accuracy of the detection unit 440 can be improved by detecting the lip motion in connection with the audio feature.

Preferably, before detecting the lip motion, the detection unit 140 can be pre-trained with the visual feature of the mouth region as extracted by the visual feature extraction unit 440 and the audio feature extracted by the audio feature extraction unit 450. The detection unit 440 can be trained by using a number of known training approaches.

Fig. 5 shows a block diagram of a device 50 for video-based lip motion detection according to another embodiment of the present invention. As shown in Fig. 5, the lip motion detection device 50 includes: a face search unit 510, a mouth region extraction unit 520, a visual feature extraction unit 530, a

- 18 -

detection unit 540 and a smoothing unit 550. Since the face search unit 510, mouth region extraction unit 520, visual feature extraction unit 530 and detection unit 540 as shown in Fig. 5 are similar to the face search unit 110, mouth region extraction unit 120, visual feature extraction unit 130 and detection unit 140 as shown in Fig. 1, only the smoothing unit 550 will be detailed in the following for simplicity.

The smoothing unit 550 temporally smoothes the detection result of the detection unit 450. Such smoothing can improve the accuracy of the detection result by utilizing the apriori knowledge that the status of the lip motion does not change repeatedly within a short period of time. For example, the smoothing unit 550 can be implemented using a median filter, preferably a five-point median filter. As an alternative, a median filter having another window size or any other filter can also be used.

It is to be understood that, as another implementation, both the audio feature extraction unit shown in Fig. 4 and the smoothing unit shown in Fig. 5 can be incorporated into the video-based lip motion detection device as shown in Fig. 1.

Fig. 6 illustrates a flowchart of a method 60 for video-based lip motion detection according to an embodiment of the present invention. The method 60 starts at step S610.

At step S620, a face is searched in an input video frame by the face search unit 110, 410 or 510. If any face is found, its

- 19 -

position will be transferred to the next step as input information. On the other hand, no further process will be taken on a video frame in which no face is found. Preferably, the face can be searched by using approaches such as Viola-Jones
5 face detection, Rowley face detection, meanshift tracking and particle filtering tracking.

At step S630, a mouth region is extracted from the searched face by the mouth region extraction unit 120, 420 or 520. For each found face, two mouth corners are searched first.
10 Then, the mouth region is determined based on the two found mouth corners. The two mouth corners can be located by the well-known Active Shape Model (ASM). Alternatively, the Active Appearance Model (AAM) and Snakes (also known as Active Contour Model) can also be used to locate the two mouth
15 corners. After the two mouth corners are located, a rectangular region can be determined, with the center of the rectangular region located at the middle of the line connecting two mouth corners and the longer edges of the rectangular region being parallel with the line connecting two mouth corners. Then, the
20 rectangular region is determined as the mouth region. The aspect ratio of the rectangular region is preferably 3:2. However, other aspect ratios are also applicable.

Alternatively, the mouth region can be of any other shape, such as ellipse, which contains the entire outer contour of the
25 lip. In addition, these shapes are not necessarily symmetric and

- 20 -

their centers are not necessarily coincident with the middle of the line connecting the mouth corners. Further, any shape having a large intersection with the outer contour of the lip can also be used as the mouth region.

5 At step S640, a visual feature of the mouth region is extracted from a spatial-temporal window which contains one or more consecutive mouth regions by the visual feature extraction unit 130, 430 or 530. In an embodiment of the present invention, the visual feature is described as a LBP-TOP
10 feature. However, it can be appreciated by those skilled in the art that the visual feature can also be described using other features, such as but not limited to the above mentioned spatial-temporal extension of gradient type feature. The step S640 can be performed by the visual feature extraction unit 130
15 as shown in Fig. 1.

Next, at step S650, the lip motion is detected based on the extracted visual feature of the mouth region by the detection unit 140 or 540. For example, a classifier can be used to distinguish two classes of mouth regions, i.e., mouth region
20 with lip motion and mouth region without lip motion, so as to obtain the detection result. Such classifier includes, but not limited to, Support Vector Machine (SVM), k-Nearest Neighbor classifier, Adaboost classifier, Neural Network classifier, Gaussian Process classifier and threshold classifier based on
25 feature similarity.

Finally, the method 60 ends at step S660.

Preferably, though not shown in Fig. 6, before the step S650 for detecting the lip motion, it is possible to pre-train with the visual feature of the mouth region as extracted in step S640.

5 In an embodiment, such training can be achieved by assigning a label to each extracted visual feature. For example, if there is a lip motion in the mouth region corresponding to a visual feature, a label of +1 is assigned to the feature, otherwise a label of -1 is assigned. Then, the pre-training can be carried out using a
10 number of known training approaches.

Additionally, though not shown in Fig. 6, an audio feature can be extracted from the input video frame by the audio feature extraction unit 450. The extracted audio feature is provided to step S650 along with the extracted visual feature
15 for the detection unit 440 to detect the lip motion. In particular, if a subject in the video is speaking and a synchronized audio is available, it is possible to use any known audio-based speech endpoint detection methods to detect the speech part and non-speech part. If a frame falls in the speech part, there is a
20 high probability of presence of lip motion in the mouth region in the frame. The detection accuracy can be improved by detecting the lip motion in connection with the audio feature. Accordingly, before detecting the lip motion, it is possible to pre-train with the visual feature and the audio feature.

25 Further, though not shown in Fig. 6, the method can

- 22 -

smooth the detection result obtained in Step S650 by the smoothing unit 550. Such smoothing can improve the accuracy of the detection result by utilizing the apriori knowledge that the status of the lip motion does not change repeatedly within a short period of time. For example, the smoothing can be implemented using a median filter, preferably a five-point median filter. As an alternative, a median filter having another window size or any other filter can also be used.

Fig. 7 shows a block diagram of a video-aided speech recognition system 70 having a device for lip motion detection according to an embodiment of the present invention. In the speech recognition system 70, a speech segment can be detected based on video, such that the accuracy of speech detection in noisy environment can be improved.

In particular, the speech recognition system 70 shown in Fig. 7 includes a microphone 710, a camera 720, a lip motion detection device 730, a speech segment detector 740, a feature extractor 750 and a speech recognizer 760. The microphone 710 and the camera 720 capture video and audio signals in real time, respectively. The speaker faces the camera 720 and the microphone 710 when speaking. The captured video is sent to the lip motion detection device 730 and the captured audio is sent to the speech segment detector 740. Herein, the lip motion detection device 730 can be implemented by the lip motion detection device 10 shown in Fig. 1, the lip motion detection

- 23 -

device 40 shown in Fig. 4 and the lip motion detection device 50 shown in Fig. 4.

If the lip motion detection device 730 detects a lip motion, it sends start time and end time of the lip motion to the speech
5 segment detector 740, which then extracts a speech segment based on the start time and end time of the lip motion as received. (a) of Fig. 8 shows the audio signal received by the microphone 710 and (b) of Fig. 8 shows the lip motion detection
10 result. It's clear that the speech endpoints and lip motion signal endpoints match with each other quite well. (c) of Fig. 8 shows the extracted speech segment based on the lip motion signal.

The extracted speech segment is then sent to the feature extractor 750 in which a audio feature is extracted and sent to
15 the speech recognizer 760 for recognizing the speech and outputting a recognition result.

Fig. 9 shows a block diagram of a video conference system 90 having a device for lip motion detection according to an embodiment of the present invention. The system is capable of
20 automatically turning on off a microphone and providing a close up for the current speaker in a case where there are multiple subjects.

Specifically, the video conference system 90 includes a microphone 910, a camera 920, a lip motion detection device
25 930, a video frame cropper 940 and a transmitter 950. The

- 24 -

microphone 910 captures audio signal in real time and the camera 920 captures video in real time. The speaker faces the camera when speaking to another party of the conference. The captured video is sent to both the lip motion detection device 5 930 and the video frame cropper 940. Herein, the lip motion detection device 930 can be implemented by the lip motion detection device 10 shown in Fig. 1, the lip motion detection device 40 shown in Fig. 4 and the lip motion detection device 50 shown in Fig. 4.

10 The video frame cropper 940 crops video from the video captured by the camera 920. At the start time of the lip motion, the lip motion detection device 930 turns on the video frame cropper 940 and sends the position of the speaker to the video frame cropper 940. If there are multiple subjects (speakers), the 15 video frame cropper crops video frame and resizes it (by means of zooming) to provide a close up for the current speaker. The cropped video and the audio from the microphone 910 are sent to the transmitter 950 for transmission. At the end time of the lip motion, the lip motion detection device 930 turns off the 20 video frame cropper 940 to stop cropping video. At this time, the transmitter 950 transmits only the video captured by the camera 920.

It is to be understood that it is not necessary for the video conference system 90 to include a video frame cropper 940. At 25 this time, at the start time of the lip motion, the lip motion

- 25 -

detection device 930 controls the transmitter 950 to transmit the audio captured by the microphone 910 and the video captured by the camera 920. At the end time of the lip motion, the lip motion detection device 930 controls the transmitter 950
5 to transmit only the video captured by the camera 920.

With the present invention, it is possible to implement subject-independent lip motion detection for a training set containing a limited number of subjects. Compared with prior art, the present invention enables a higher detection rate for a
10 subject not contained in the training set. According to the present invention, the retraining or adaptation for different users for improving the detection rate is not necessary any more, and thus the usability is improved.

The present invention has been described above with
15 reference to the preferred embodiments thereof. It should be understood that various modifications, alternations and variations can be made by those skilled in the art without departing from the spirits and scope of the present invention. Therefore, the scope of the present invention is not limited to
20 the above particular embodiments but only defined by the claims as attached and their equivalents.

- 26 -

CLAIMS

1. A device for video-based lip motion detection, comprising:

5 -a face search unit adapted for searching a face in an input video frame;

 -a mouth region extraction unit adapted for extracting a mouth region from the searched face;

 -a visual feature extraction unit adapted for extracting a
10 visual feature of the mouth region; and

 -a detection unit adapted for detecting lip motion based on the extracted visual feature of the mouth region.

2. The device for video-based lip motion detection
15 according to claim 1, wherein the visual feature extraction unit is adapted for extracting at least one of a gradient on spatial-temporal plane and a Local Binary Pattern (LBP) code on spatial-temporal plane for each pixel in a spatial-temporal window and extracting a visual feature of the mouth region
20 based on the extraction result.

3. The device for video-based lip motion detection according to claim 1, wherein the detection unit is pre-trained with the extracted visual feature of the mouth region.

- 27 -

4. The device for video-based lip motion detection according to claim 1, further comprising:

-a smoothing unit adapted for smoothing the detection result of the detection unit.

5

5. The device for video-based lip motion detection according to claim 1, further comprising:

-an audio feature extraction unit adapted for extracting an audio feature corresponding to the input video frame;

10 wherein the detection unit is adapted for detecting lip motion based on the visual feature extracted by the visual feature extraction unit and the audio feature extracted by the audio feature extraction unit.

15 6. The device for video-based lip motion detection according to claim 5, wherein the detection unit is pre-trained with the extracted visual feature and audio feature.

7. The device for video-based lip motion detection
20 according to claim 1, wherein the visual feature comprises a Local Binary Pattern on Three Orthogonal Planes (LBP-TOP)-based visual feature.

8. The device for video-based lip motion detection
25 according to claim 1, wherein the mouth region is a rectangle,

- 28 -

the center of the rectangle being located at the middle of the line connecting two mouth corners and the longer edges of the rectangle being parallel with the line connecting two mouth corners.

5

9. The device for video-based lip motion detection according to claim 1, wherein the detection unit comprises a Support Vector Machine (SVM).

10

10. The device for video-based lip motion detection according to claim 4, wherein the smoothing unit comprises a median filter.

15

11. The device for video-based lip motion detection according to claim 1, wherein the face search unit comprises a Viola-Jones face detector.

20

12. The device for video-based lip motion detection according to claim 1, wherein the mouth region extraction unit is adapted for extracting the mouth region from the searched face by using an Active Shape Model (ASM).

25

13. The device for video-based lip motion detection according to claim 1, wherein the visual feature extraction unit is further adapted for extracting at least one of a gradient on

- 29 -

image plane and a Local Binary Pattern (LBP) code on image plane for each pixel in a spatial-temporal window and extracting a visual feature of the mouth region based on the overall extraction result

5

14. A method for video-based lip motion detection, comprising the following steps of:

- searching a face in an input video frame;
- extracting a mouth region from the searched face;
- 10 -extracting a visual feature of the mouth region; and
- detecting lip motion based on the extracted visual feature of the mouth region.

15 15. The method for video-based lip motion detection according to claim 14, further comprising:

- extracting at least one of a gradient on spatial-temporal plane and a Local Binary Pattern (LBP) code on spatial-temporal plane for each pixel in a spatial-temporal window, prior to extracting a visual feature of the mouth region;
- 20 wherein the visual feature of the mouth region is extracted based on the extraction result.

16. The method for video-based lip motion detection according to claim 14, further comprising:

- 25 -pre-training with the extracted visual feature of the

- 30 -

mouth region, prior to detecting lip motion.

17. The method for video-based lip motion detection according to claim 14, further comprising:

5 -smoothing the detection result.

18. The method for video-based lip motion detection according to claim 14, further comprising:

 -extracting an audio feature corresponding to the input
10 video frame;

 wherein the lip motion is detected based on the extracted visual feature and audio feature.

19. The method for video-based lip motion detection
15 according to claim 18, further comprising:

 -pre-training with the extracted visual feature and audio feature, prior to detecting lip motion.

20. The method for video-based lip motion detection
20 according to claim 14, wherein the visual feature comprises a Local Binary Pattern on Three Orthogonal Planes (LBP-TOP)-based visual feature.

21. The method for video-based lip motion detection
25 according to claim 14, wherein the mouth region is a rectangle,

- 31 -

the center of the rectangle being located at the middle of the line connecting two mouth corners and the longer edges of the rectangle being parallel with the line connecting two mouth corners.

5

22. The method for video-based lip motion detection according to claim 14, wherein the lip mouth is detected by using a Support Vector Machine (SVM).

10

23. The method for video-based lip motion detection according to claim 17, wherein the detection result is smoothed by using a median filter.

24. The method for video-based lip motion detection
15 according to claim 14, wherein the face is searched from the input video frame by using a Viola-Jones face detector.

25. The method for video-based lip motion detection according to claim 14, wherein the mouth region is extracted
20 from the searched face by using an Active Shape Model (ASM).

26. The method for video-based lip motion detection according to claim 14, further comprising:

-extracting at least one of a gradient on image plane and a
25 Local Binary Pattern (LBP) code on image plane for each pixel in

- 32 -

a spatial-temporal window and extracting a visual feature of the mouth region based on the overall extraction result.

27. A speech recognition system, comprising:

- 5 -a microphone adapted for capturing an audio signal;
- a camera adapted for capturing a video signal;
- a device for video-based lip motion detection according to any one of claims 1-13, adapted for detecting lip motion based on the video signal captured by the camera to obtain start time
- 10 and end time of the lip motion;
- a speech segment detector adapted for extracting a speech segment based on the audio signal captured by the microphone and the start time and the end time of the lip motion;
- 15 -a feature extractor adapted for extracting an audio feature from the extracted speech segment; and
- a speech recognizer adapted for recognizing speech based on the extracted audio feature.

20 28. A video conference system, comprising:

- a microphone adapted for capturing an audio signal;
- a camera adapted for capturing a video signal;
- a device for video-based lip motion detection according to any one of claims 1-13, adapted for detecting lip motion based
- 25 on the video signal captured by the camera to obtain start time

- 33 -

and end time of the lip motion; and

-a transmitter;

wherein the device for video-based lip motion detection is adapted for controlling the transmitter to transmit the audio
5 signal captured by the microphone and the video signal captured by the camera at the start time of the lip motion and for controlling the transmitter to transmit only the video signal captured by the camera at the end time of the lip motion.

10 29. The video conference system according to claim 28, further comprising:

-a video frame cropper adapted for cropping video from the video signal captured by the camera;

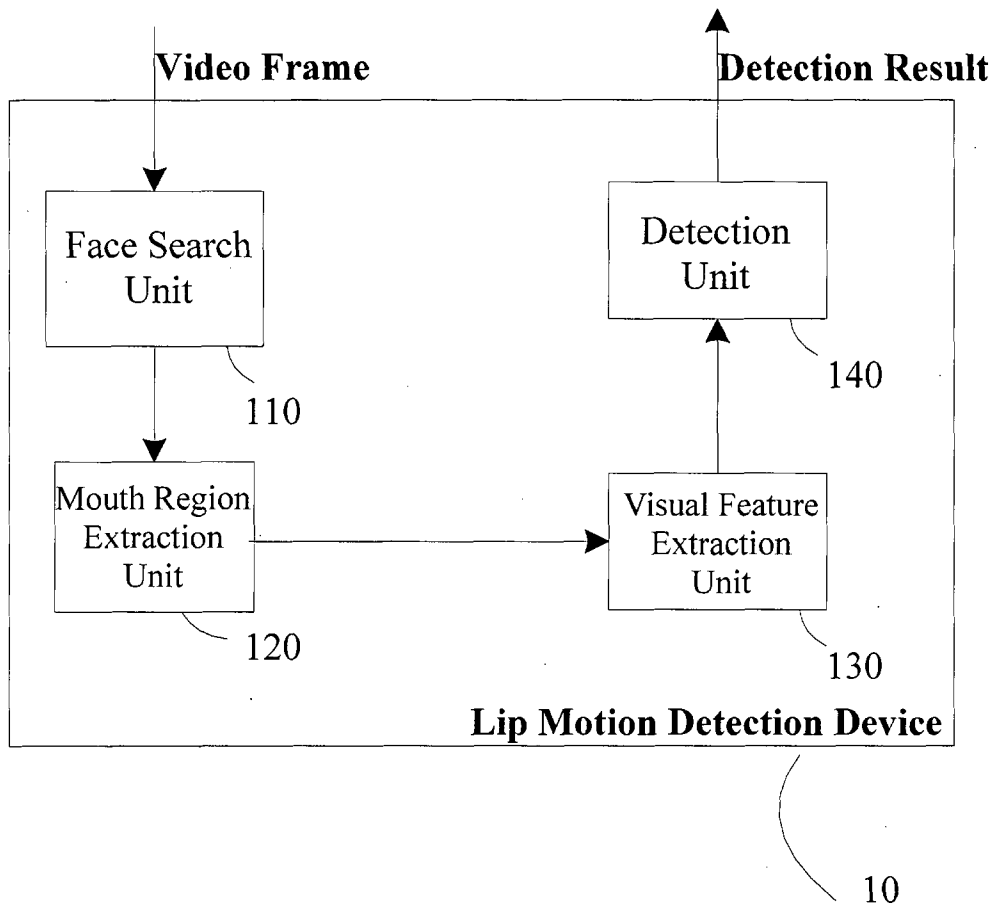
wherein the device for video-based lip motion detection is
15 adapted for enabling the video frame cropper and controlling the transmitter to transmit the video signal captured by the microphone and the video cropped by the video frame cropper at the start time of the lip motion and for disabling the video frame cropper and controlling the transmitter to transmit only the
20 video signal captured by the camera at the end time of the lip motion.

30. The video conference system according to claim 29, wherein the video frame cropper is adapted for cropping a
25 close-up of a speaker who is currently speaking by means of

zooming.

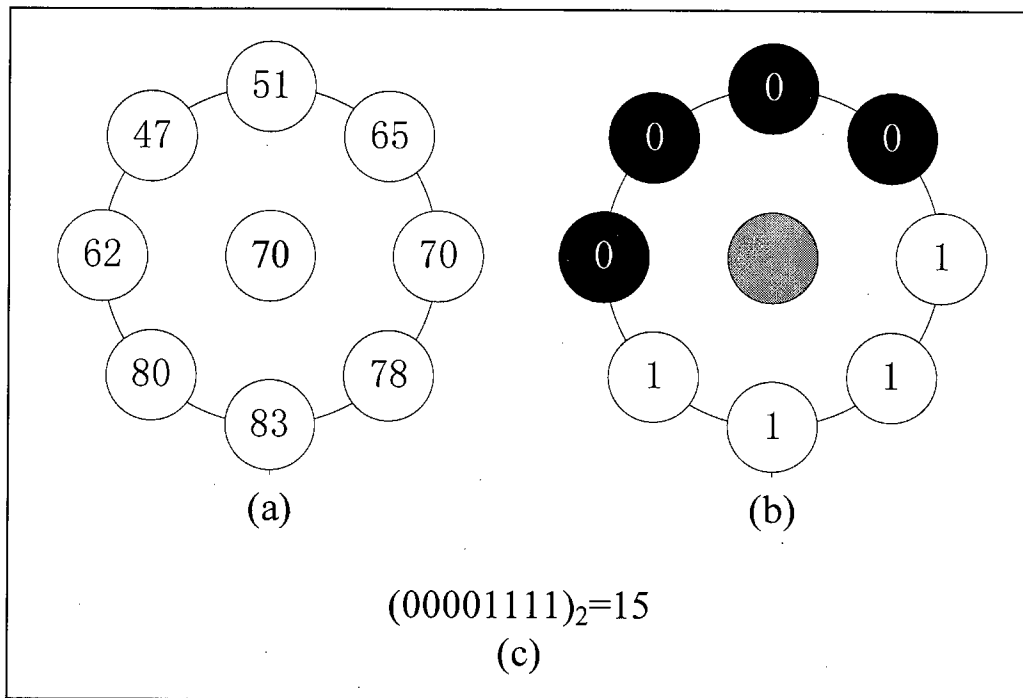
1/9

FIG. 1



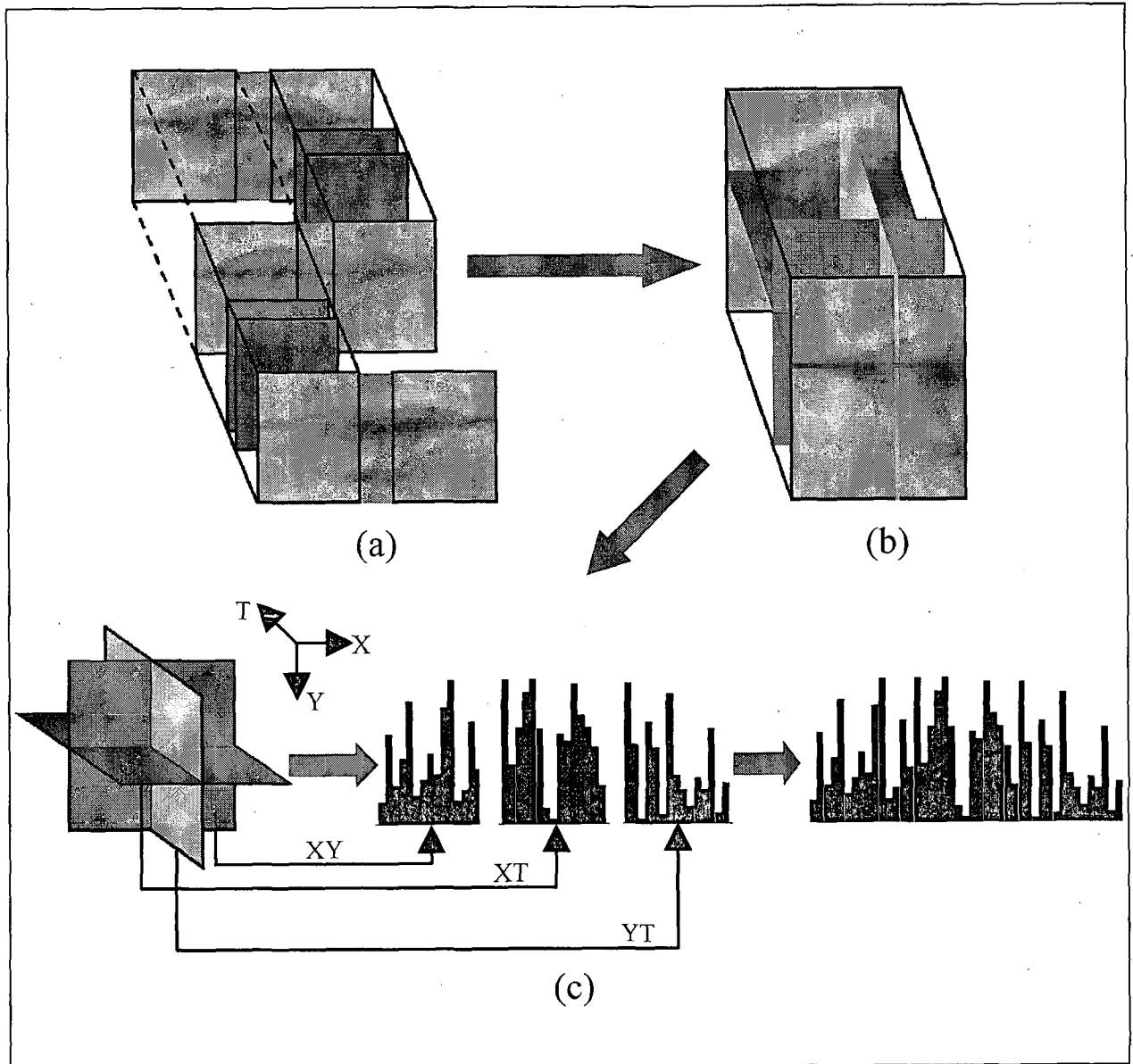
2/9

FIG. 2



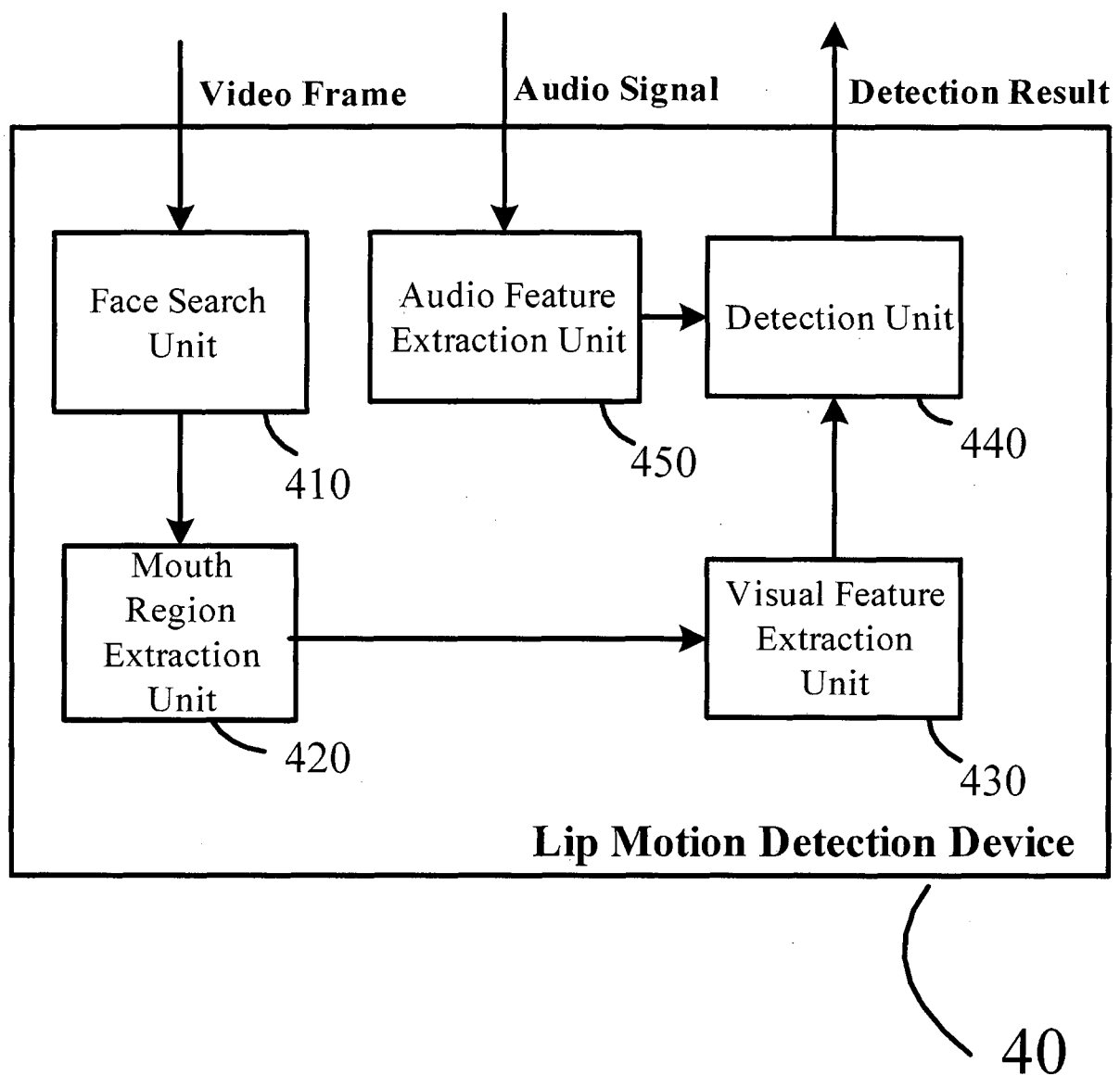
3/9

FIG. 3



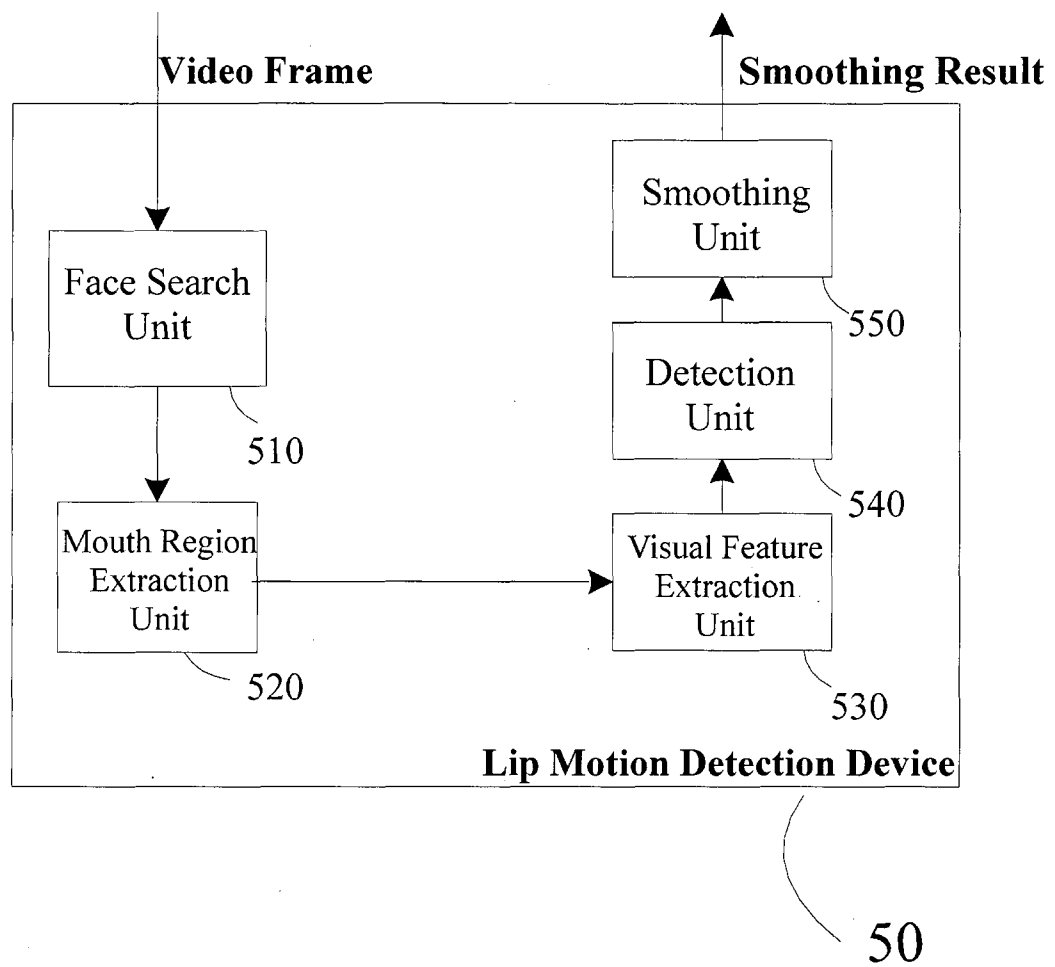
4/9

FIG. 4



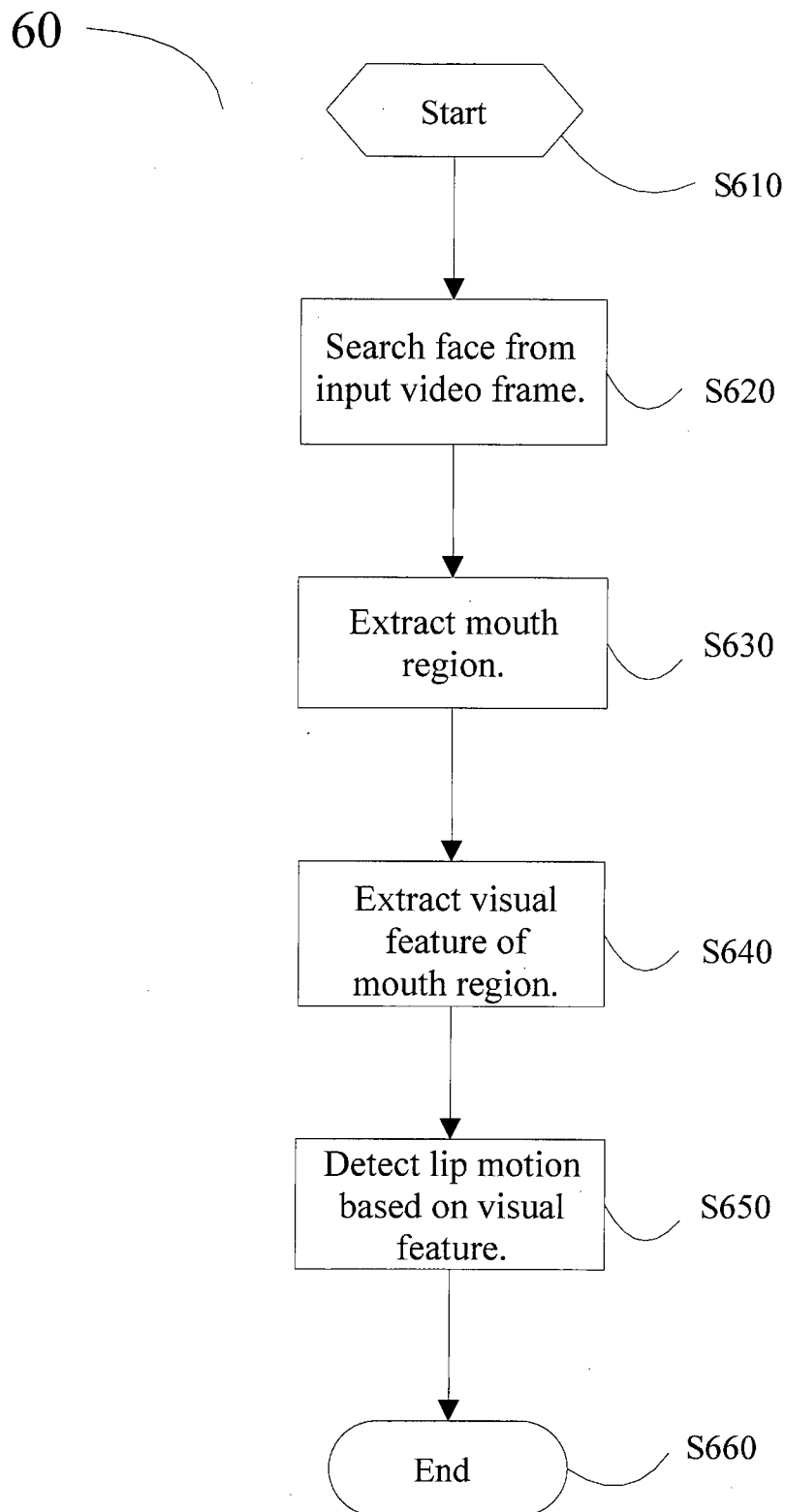
5/9

FIG. 5



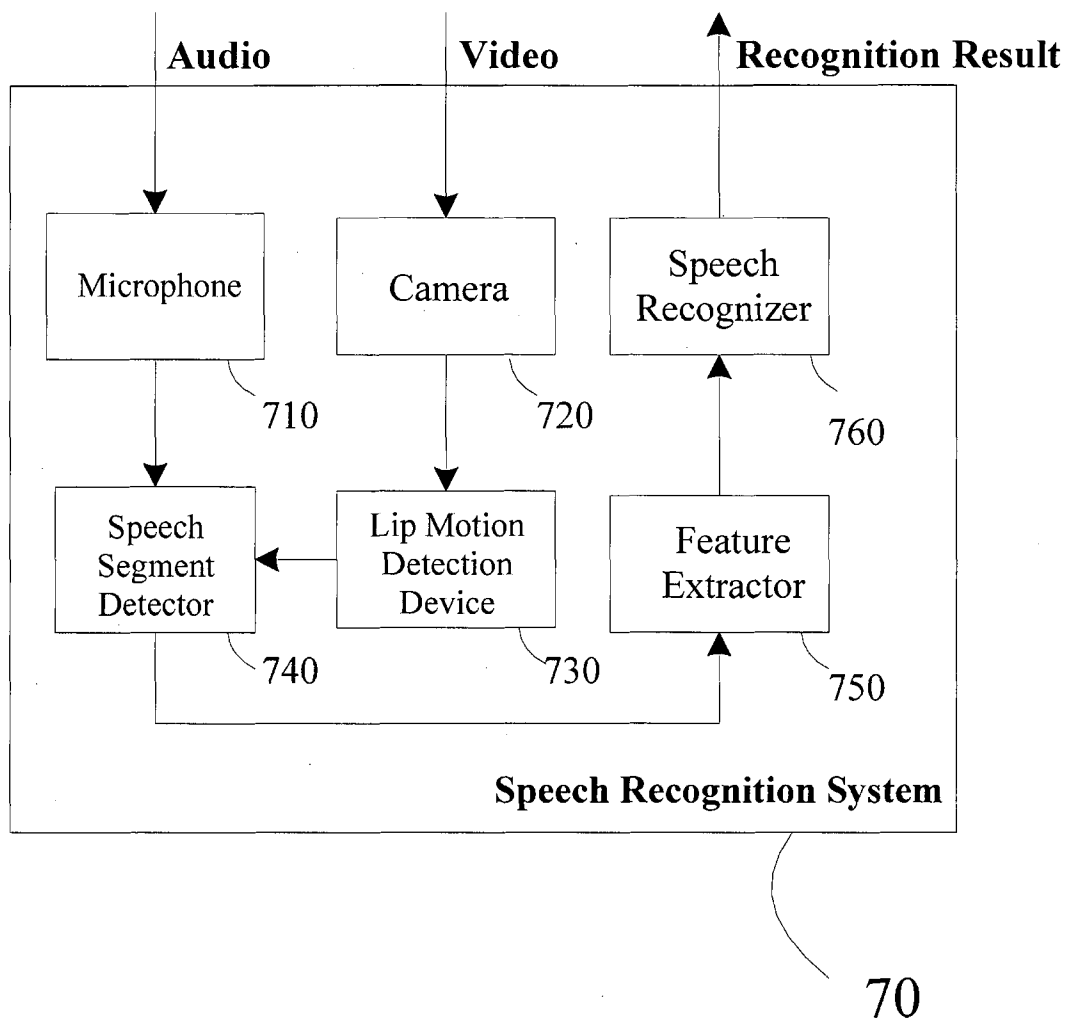
6/9

FIG. 6

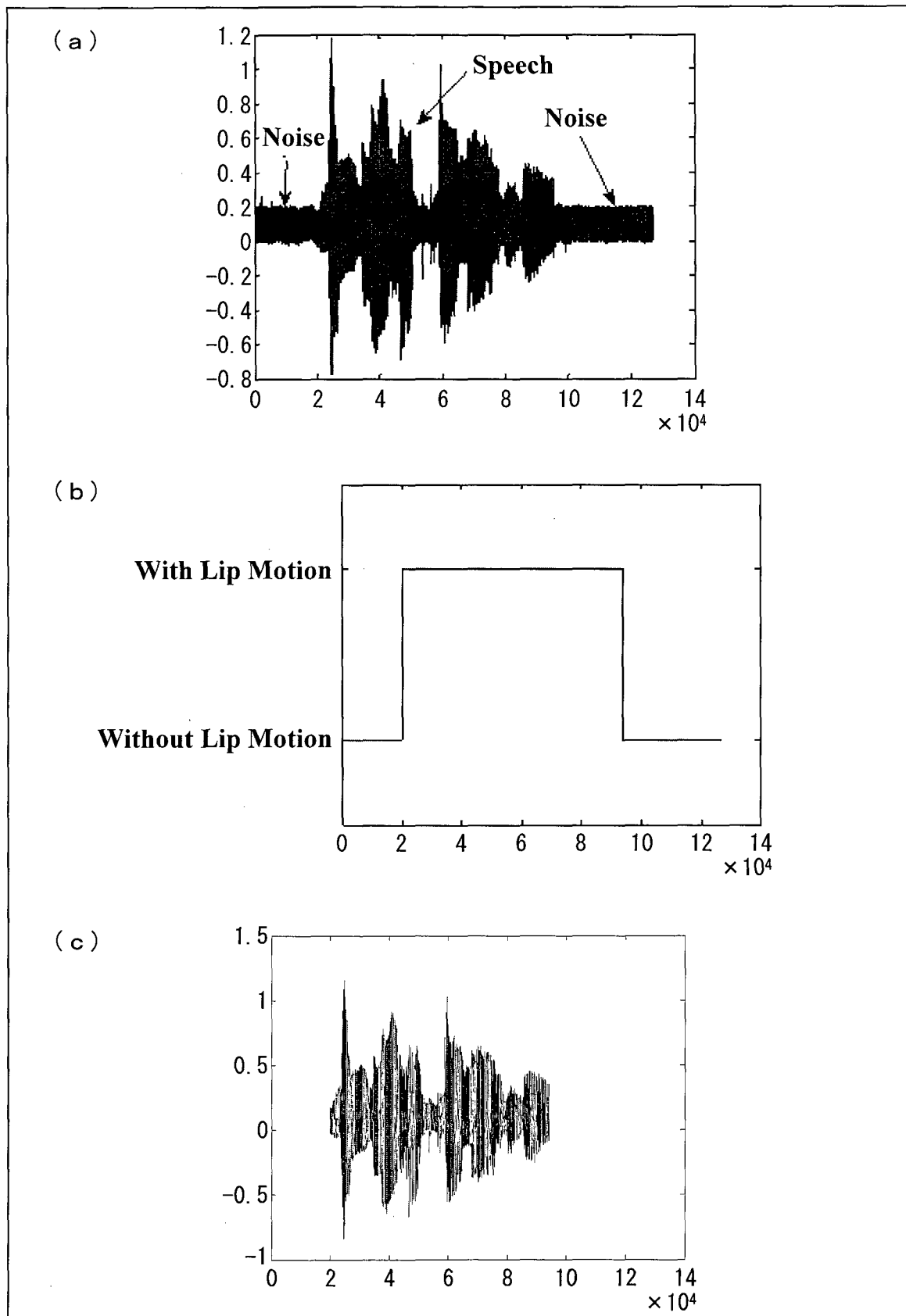


7/9

FIG. 7

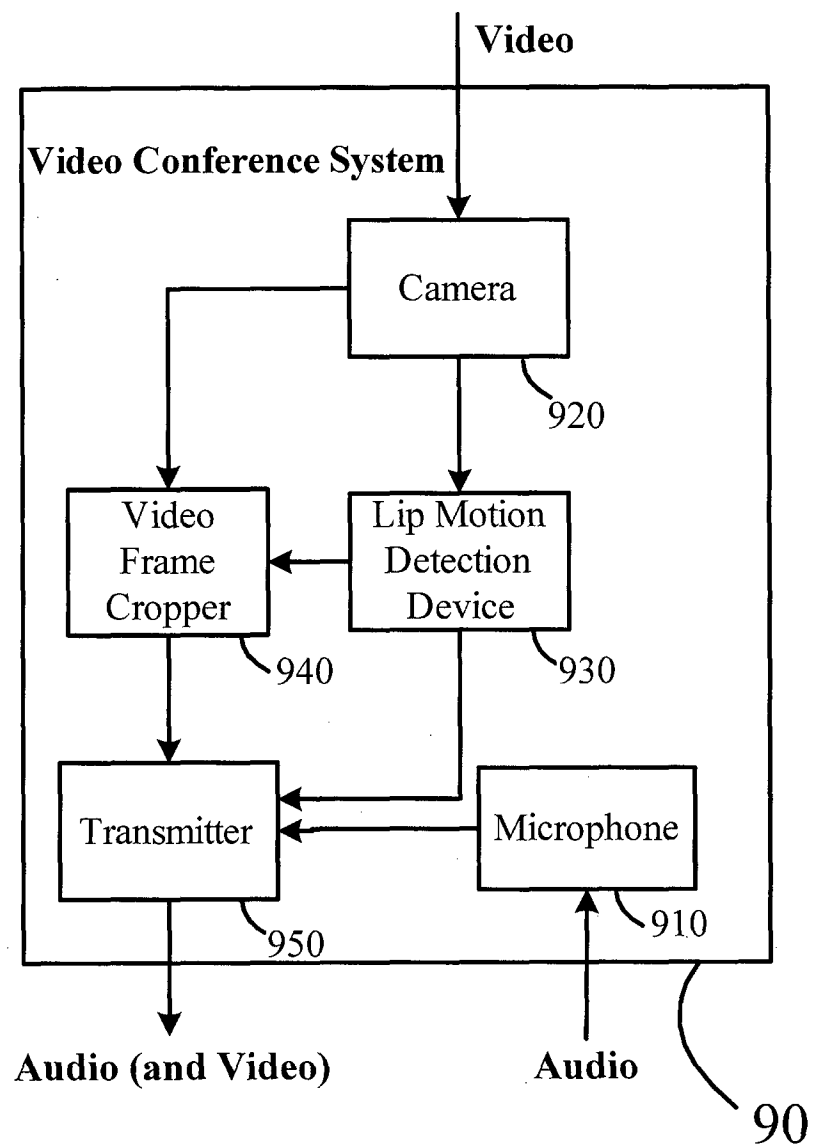


8/9

FIG. 8

9/9

FIG. 9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/057677

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06T7/20 (2006.01) i, G06T7/00 (2006.01) i

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)

Int.Cl. G06T7/20, G06T7/00

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

Published examined utility model applications of Japan 1922-1996
 Published unexamined utility model applications of Japan 1971-2012
 Registered utility model specifications of Japan 1996-2012
 Published registered utility model applications of Japan 1994-2012

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-240154 A (Hitachi, Ltd.) 2004.08.26, [claim 1] (No Family)	1-6, 8-19, 21-27
A		7, 20, 28-30
Y	JP 7-225841 A (SHARP CORPORATION) 1995.08.22, [claim 2] (No Family)	1-6, 8-19, 21-27
Y	Atsushi SAYO et al., A Study on Classifier Generation Methods for Personal Authentication System Using Lip Variation, IEICE Technical Report, Vol.110 No.217, The Institute of Electronics, Information and Communication Engineers, 2010.09.28, pp.7-12	2, 13, 15, 26
A		7, 20



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

17.04.2012

Date of mailing of the international search report

24.04.2012

Name and mailing address of the ISA/JP

Japan Patent Office

3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan

Authorized officer

Sato Minoru

Telephone No. +81-3-3581-1101 Ext. 3531

5H

3247

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/057677

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2009-98901 A (Nippon Telegraph and Telephone Corporation) 2009.05.07, [0037] (No Family)	4, 10, 17, 23
Y	JP 2010-204984 A (NISSAN MOTOR CO., LTD.) 2010.09.16, [0031] (No Family)	9, 22
Y	P.Viola, M.Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp.511-518	11, 24
Y	JP 2006-202276 A (Fuji Photo Film. Co. Ltd.) 2006.08.03, [0003] & US 2006/0133672 A1	12, 25
Y	JP 2003-195883 A (TOSHIBA CORPORATION) 2003.07.09, [0019]-[0021] (No Family)	27