Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 3024565 C 2022/09/06

(11)(21) 3 024 565

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2017/04/12
(87) Date publication PCT/PCT Publication Date: 2017/11/23
(45) Date de délivrance/Issue Date: 2022/09/06
(85) Entrée phase nationale/National Entry: 2018/11/16
(86) N° demande PCT/PCT Application No.: CN 2017/080196
(87) N° publication PCT/PCT Publication No.: 2017/198014
(30) Priorité/Priority: 2016/05/19 (CN201610340549.4)

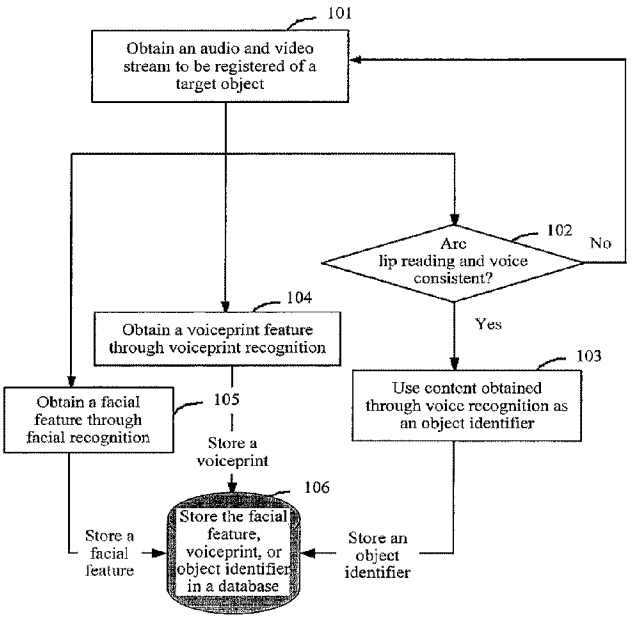
(51) Cl.Int./Int.Cl. *G06F 21/32* (2013.01)

(72) Inventeurs/Inventors:
LI, PENG, CN;
SUN, YIPENG, CN;
XIE, YONGXIANG, CN;
LI, LIANG, CN

(73) Propriétaire/Owner:
ADVANCED NEW TECHNOLOGIES CO., LTD., KY

(74) Agent: KIRBY EADES GALE BAKER

(54) Titre : PROCEDE ET APPAREIL D'AUTHENTIFICATION D'IDENTITE
(54) Title: IDENTITY AUTHENTICATION METHOD AND APPARATUS



(57) Abrégé/Abstract:
The present application provides an identity authentication method and apparatus. The method includes obtaining a collected audio and video stream generated by a target object to be authenticated; determining whether lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object; obtaining a model physiological feature corresponding to the object identifier from object registration information, if the pre-stored object registration information includes the object identifier; performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object; and comparing the physiological feature of the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication condition, determining that the target object has been authenticated. The present application improves the efficiency and reliability of identity authentication.

ABSTRACT

The present application provides an identity authentication method and apparatus. The method includes obtaining a collected audio and video stream generated by a target object to be authenticated; determining whether lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object; obtaining a model physiological feature corresponding to the object identifier from object registration information, if the pre-stored object registration information includes the object identifier; performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object; and comparing the physiological feature of the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication condition, determining that the target object has been authenticated. The present application improves the efficiency and reliability of identity authentication.

IDENTITY AUTHENTICATION METHOD AND APPARATUS

TECHNICAL FIELD

[0001] The present application relates to network technologies, and in particular, to an identity authentication method and apparatus.

BACKGROUND

[0002] With the development of the Internet technologies, network-based services are more widely used by people. For example, people can send and receive emails by using a mailbox, shop online, or even work online. Some applications have high security requirements, and a user's identity needs to be authenticated. For example, a user's identity needs to be authenticated before payment is authorized for online shopping; or the user can log in to an application with a relatively high security requirement after the user's identity has been authenticated. In related technologies, there are several identity authentication methods used by the Internet, such as facial recognition and voiceprint recognition. However, these commonly used authentication methods are relatively complicated. For example, a user needs to enter the user's ID and then verify a voiceprint for voiceprint recognition. In addition, existing authentication methods are of relatively low reliability. For example, an attacker can achieve facial recognition by using an analog video or recording. Even if verification is performed based on two combined authentication methods, for example, with reference to a face and a voiceprint, as these authentication methods are relatively independent from each other, an attacker can break through the authentication processes. The previously described disadvantages of the identity authentication methods can create some risks to application security.

SUMMARY

[0003] In view of the above, the present application provides an identity

authentication method and apparatus, so as to improve the efficiency and reliability of identity authentication.

[0003a] Certain exemplary embodiments can provide an identity authentication method, wherein the method comprises: obtaining an audio and video stream, wherein the audio and video stream is generated by a target object to be authenticated; determining whether lip reading and voice in the audio and video stream are consistent; in response to determining that the lip reading and the voice are consistent, using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object; determining whether pre-stored object registration information comprises the object identifier; in response to determining that the pre-stored object registration information comprises the object identifier, obtaining a model physiological feature corresponding to the object identifier from the pre-stored object registration information; performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object; comparing the physiological feature of the target object with the model physiological feature to obtain a comparison result; determining whether the comparison result satisfies an authentication condition; and in response to determining that the comparison result satisfies an authentication condition, determining that the target object has been authenticated.

[0004] Specifically, the present application is implemented by using the following technical solutions.

[0005] According to another aspect, an identity authentication method is provided, and the method includes obtaining a collected audio and video stream, where the audio and video stream is generated by a target object to be authenticated; determining whether the lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, using voice content obtained by performing voice recognition on an audio stream of the audio and video stream as an object identifier of the target object; obtaining a model physiological feature corresponding to the object identifier from object registration information, if the pre-stored object registration information includes the object identifier; performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object; and comparing the physiological feature of

the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication condition, determining that the target object has been authenticated.

[0006] According to another aspect, an identity authentication apparatus is provided, and the apparatus includes an information acquisition module, configured to obtain a collected audio and video stream, where the audio and video stream is generated by a target object to be authenticated; an identifier determining module, configured to determine whether the lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, use voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object; an information management module, configured to obtain a model physiological feature corresponding to the object identifier from object registration information, if the pre-stored object registration information includes the object identifier; a feature recognition module, configured to perform physiological recognition on the audio and video stream to obtain a physiological feature of the target object; and an authentication processing module, configured to compare the physiological feature of the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication condition, determine that the target object has been

authenticated.

[0007] According to the identity authentication method and apparatus provided in the present application, a user identifier is obtained through recognition of an audio and video stream during authentication of a user, and a facial feature and a voiceprint feature can also be verified by using the same audio and video stream. It simplifies the user's operation, improves the authentication efficiency, maintains a one-to-one authentication model, and ensures recognition accuracy. In addition, in the method, consistency between lip reading and voice is determined to ensure that a target object is a living object instead of a fake video record of an attacker, thereby improving the authentication security and reliability.

BRIEF DESCRIPTION OF DRAWINGS

[0008] FIG. 1 illustrates an identity registration procedure, according to an example implementation of the present application;

[0009] FIG. 2 illustrates a procedure of determining consistency between lip reading and voice, according to an example implementation of the present application;

[0010] FIG. 3 illustrates a facial feature recognition procedure, according to an example implementation of the present application;

[0011] FIG. 4 illustrates a voiceprint feature recognition procedure, according to an example implementation of the present application;

[0012] FIG. 5 illustrates an identity authentication procedure, according to an example implementation of the present application;

[0013] FIG. 6 is a structural diagram of an identity authentication apparatus, according to an example implementation of the present application; and

[0014] FIG. 7 is a structural diagram of an identity authentication apparatus, according to an example implementation of the present application.

DESCRIPTION OF EMBODIMENTS

[0015] Example implementations are described in detail here, and examples of the example implementations are presented in the accompanying drawings. When the following description relates to the accompanying drawings, unless specified otherwise, same numbers in different accompanying drawings represent same or

similar elements. Implementations described in the following example implementations do not represent all implementations consistent with the present application. Instead, they are only examples of apparatuses and methods consistent with some aspects of the present application that are described in detail in the appended claims.

[0016] An identity authentication method provided in an implementation of the present application can be applied to Internet identity authentication. For example, a user can log in to a network application after having been authenticated by using the method, so as to ensure the security of application.

[0017] The following uses an application with a relatively high security requirement as an example. Assume that the application can run on a user's intelligent device, for example, a smartphone or an intelligent tablet. When the user needs to log in to the application on the intelligent device, an audio and video stream can be collected by using a camera and a microphone on the intelligent device. For example, the user can read an application ID of the user to a camera and a microphone of a mobile phone. The application ID can be an account number "123456" that is registered by the user in the application. After the user reads the application ID, the mobile phone can collect an audio and video stream of the user, including a video image and reading voice.

[0018] According to the identity authentication method in this implementation of the present application, the collected audio and video stream can be processed. Before authentication, the user needs to perform an identity registration procedure to perform the identity authentication later. The registration procedure is also performed based on the collected audio and video stream. The following separately describes an identity registration procedure and an identity authentication procedure. In addition, this implementation does not impose a limitation on a device that performs the identity registration or authentication processing in actual applications. For example, after the audio and video stream of the user is collected, the smartphone can transmit the audio and video stream to a backend server of the application, or a part of the processing can be performed on a client software side of the smartphone, and the other parts of the processing can be performed on a server side; or other methods can be used.

Identity registration

[0019] In the method in this implementation, when the user performs identity registration, two types of information can be included. One type of information is an object identifier. For example, when a user logs in to an application, the user can be referred to as a target object. When the user registers with the application, information used to distinguish the user from another user is an object identifier in the application, for example, can be an account number 123456 of the user in an application, and the account number 123456 is an object identifier of the target object. The other type of information is physiological information that can uniquely identify a user, for example, a voiceprint feature of the user or a facial feature of the user. Usually, voiceprints and faces of different people are different, and physiological information that identifies each user can be referred to as a model physiological feature.

[0020] A mapping relationship is established between the two types of information: the object identifier and the model physiological feature, and the mapping relationship is stored. The correspondingly stored object identifier and model physiological feature of the target object can be referred to as "object registration information". For example, user Xiao Zhang can store object registration information of the user as "123456—model physiological feature A". For a more accurate identification of the user, the physiological information included in the model physiological feature used in this example can be at least two types of physiological information, for example, a face and a voiceprint.

[0021] FIG. 1 illustrates an example identity registration procedure. The procedure includes the following processing:

[0022] Step 101: Obtain an audio and video stream to be registered of a target object.

[0023] For example, a user registers with an application, and the user can read an account number "123456" of the application to an intelligent device of the user, such as a mobile phone. In this example, the user that is performing registration can be referred to as a target object; and a camera and a microphone of the intelligent device can collect an audio and video stream generated when the user reads the account number. The audio and video stream collected during registration can be referred to as an audio and video stream to be registered, including an audio stream and a video stream. The audio stream is reading voice of the user, and the video stream is a

reading video image of the user.

[0024] After the audio and video stream in the present step is obtained, the three following aspects of processing can be performed to complete a user's registration. For details, references can be still made to FIG. 1.

[0025] In one aspect, processing is as follows: Step 102: Determine whether the lip reading and voice in the audio and video stream to be registered are consistent.

[0026] The consistency here indicates a mapping relationship between a lip movement and a movement indicated by voice. For example, assume that voice is "the weather today is sunny", where the voice is generated at a low speed, but a lip movement corresponds to "the weather today is sunny" generated at a high speed. It is obvious that the voice and the lip movement do not correspond to each other: when the lip movement has stopped (content is completely read), the voice continues (...today is sunny). This may occur in many possible situations, such as when an attacker attempts to pass a users' ID and facial detection, the attacker can attack facial detection by using a previous video record of the user (the attacked user), and the attacker reads the user's ID to attack recognition of the voice content ID. As such, the attacker may be authenticated. However, in such an attack, lip reading and voice are usually inconsistent, and it can be determined that the reader is not the user.

[0027] As shown in step 102, if it is determined that the lip reading and the voice in the audio and video stream to be registered are inconsistent, a registration failure can be notified to the user. Alternatively, as shown in FIG. 1, go to step 101 to collect an audio and video stream again to prevent a processing error.

[0028] Otherwise, if it is determined that the lip reading and the voice in the audio and video stream to be registered are consistent, perform step 103: Use voice content obtained by performing voice recognition on the audio stream in the collected audio and video stream, as an object identifier of the target object. Voice recognition uses computer technologies to automatically recognize voice content of a person, that is, the recognition process of converting voice to content. For example, after voice recognition is performed on audio of reading "123456" by the user that is to perform registration, obtained voice content of the audio stream is "123456", and content obtained through recognition can be used as the user's identifier, namely, the user's ID.

[0029] The previous voice recognition of the audio stream can be performed after it is determined that the lip reading and the voice are consistent, to obtain the object identifier, or can be performed in a process of determining whether the lip reading and

the voice are consistent, to obtain the object identifier.

[0030] In another aspect, physiological recognition is performed on the audio and video stream to be registered, to obtain a model physiological feature of the audio and video stream to be registered. In this example, a physiological feature includes a facial feature and a voiceprint feature, but is not limited to the two features, provided that an example feature can uniquely identify a user, and can be used to distinguish between physiological features of different users. In this aspect, as shown in step 104, voiceprint recognition can be performed on the audio stream in the audio and video stream to be registered, to obtain a voiceprint feature of the target object.

[0031] In yet another aspect, facial detection is performed on the video stream in the audio and video stream to be registered, to obtain a facial feature of the target object.

[0032] In the registration procedure, the detected facial feature can be referred to as a model facial feature and used as a criterion in a subsequent authentication process. Likewise, the detected voiceprint feature can also be referred to as a model voiceprint feature, and the model voiceprint feature and the model facial feature can be collectively referred to as a model physiological feature.

[0033] In this implementation, the model physiological feature and the object identifier of the target object are also referred to as object registration information. After it is determined that data in the object registration information is complete, in step 106, the object identifier of the target object and the corresponding model physiological feature are stored in a database as the object registration information.

[0034] In addition, an execution order of the three aspects shown in FIG. 1 is not limited. For example, after the audio and video stream to be registered is obtained in step 101, the three aspects can be executed in parallel. If the lip reading and the voice are inconsistent, a recognized voiceprint feature and a recognized facial feature may not be stored. Alternatively, after it is determined that the lip reading and the voice are consistent, detection and recognition are then performed on the voiceprint feature and the facial feature.

[0035] FIG. 2 illustrates a procedure of determining consistency between lip reading and voice in FIG. 1. The procedure can include the following steps:

[0036] Step 201: Perform endpoint detection based on an audio stream in an audio and video stream to be registered. In the present step, a start time and an end time of an audio stream can be detected in the continuous audio stream.

[0037] Step 202: Extract continuous voice features based on the audio stream, where the features include but are not limited to an MFCC feature and an LPCC feature. The features extracted in the present step can be used for voice recognition.

[0038] Step 203: Identify a voice syllable and a corresponding time point in the audio stream. In the present step, each syllable in the audio stream can be identified based on the voice features extracted in step 202, and a corresponding appearing time point and a corresponding disappearing time point of the syllable can be determined. The voice recognition method includes but is not limited to methods such as a hidden Markov model (HMM), a deep neural network (DNN), and a Long Short Time Model (LSTM).

[0039] Step 204: Detect locations of lips based on a video stream in the audio and video stream to be registered. In the present step, the locations of the lips can be detected from a video image.

[0040] Step 205: Determine quality of a detected lip image. For example, parameters such as definition and exposure of the locations of the lips can be determined. If the definition is low or the degree of exposure is too high, it is determined that the image is not qualified; and in this case, collect an audio and video stream to be registered again. If the image is qualified, proceed to step 206 to perform lip reading recognition.

[0041] Step 206: Extract continuous features of the lips. In the present step, a feature can be extracted from continuous lip images, and the feature includes but is not limited to a local image descriptor such as a bare pixel, an LBP, Gabor, SIFT, or Surf.

[0042] Step 207: Identify a lip reading syllable and a corresponding time point in the video stream. In the present step, the lip reading syllable can be identified by using methods such as a hidden Markov model (HMM) and a Long Short Time Model. A time point corresponding to a lip reading syllable in a video time sequence is also determined by using the model during lip reading recognition.

[0043] Step 208: Determine whether both the lip reading syllable and the voice syllable are consistent with the corresponding time point. For example, in the present step, the time point information of the voice syllable can be compared with the time point information of the lip reading syllable. If a comparison result indicates consistency, it is considered that the audio stream is generated by a real person; in this case, proceed to step 209. If a comparison result indicates inconsistency, it is

suspected that the audio stream is an attack behavior; and in this case, go back to the registration procedure. In this implementation, a method of detecting consistency between both the lip reading syllable and the voice syllable and the corresponding time point is more detailed, and therefore there is higher accuracy in determining a real person's voice.

[0044] Step 209: Perform voice recognition on the voice features extracted in step 202, to obtain a user's ID, namely, the object identifier. The voice recognition method includes but is not limited to methods such as a hidden Markov model (HMM), a deep neural network (DNN), and a Long Short Time Model (LSTM).

[0045] In addition, in the example shown in FIG. 2, voice recognition of the audio stream can be performed in step 209 after it is determined that the lip reading and the voice are consistent. Alternatively, when the syllable time point in the audio stream is identified in step 203, the user's ID is obtained by performing voice recognition based on the voice feature. In this case, after it is determined that the lip reading and the voice are consistent in step 208, the user's ID obtained through recognition can be directly used as the object identifier.

[0046] FIG. 3 illustrates a facial feature recognition procedure in FIG. 1. The procedure can include the following steps:

[0047] Step 301: Detect face images based on a video stream in an audio and video stream to be registered. In the present step, a video frame image can be extracted from the video stream in the audio and video stream, and it can be detected whether a face appears in the video frame image. If yes, proceed to step 302. Otherwise, go back to the determining procedure.

[0048] Step 302: Detect quality of the face images. In the present step, facial feature point detection can be performed on the face detected in step 301, and angles of the face in both a horizontal direction and a vertical direction can be determined based on a result of the feature point detection. If both the angles do not exceed certain tilt angles, a quality requirement is satisfied. Otherwise, a quality requirement is not satisfied. In addition, definition, exposure, etc. of a face area are determined, which also need to be within certain thresholds. If the face images are of good quality, a facial feature can be better recognized.

[0049] Step 303: For face images that satisfy the quality requirement, extract eigenvectors from the face images, where the eigenvectors include but are not limited to a local binary pattern (LBP), Gabor, a convolutional neural network (CNN), etc.

[0050] Step 304: Combine a plurality of facial eigenvectors extracted in step 303, to form a unique facial feature of a user, namely, a model facial feature.

[0051] FIG. 4 illustrates a voiceprint feature recognition procedure in FIG. 1. The procedure can include the following steps.

[0052] Step 401: Obtain an audio stream in an audio and video stream to be registered.

[0053] In this example, voiceprint feature recognition can be performed based on the audio stream in the audio and video stream to be registered.

[0054] Step 402: Determine that audio quality of the audio stream satisfies a quality criterion.

[0055] In the present step, audio quality can be determined. Better quality of the collected audio stream leads to a better effect of performing voiceprint recognition on the audio. Therefore, before subsequent voiceprint recognition is performed, the quality of the audio stream can be determined first. For example, information such as signal strength and a signal-to-noise ratio of voice in the audio stream can be calculated to determine whether the voice satisfies a quality criterion. For example, the quality criterion can be that a signal-to-noise ratio falls within a certain range, or can be that signal strength of voice is greater than a strength threshold. If the audio stream is qualified, proceed to step 403. Otherwise, collect an audio and video stream to be registered again.

[0056] Step 403: Extract voiceprint eigenvectors from the audio stream.

[0057] In this example, there can be a plurality of audio and video streams to be registered. For example, the user can read the user' ID twice, and correspondingly two audio and video streams are collected. In the present step, a voiceprint eigenvector of an audio stream in each audio and video stream can be extracted. The eigenvector can be extracted in a plurality of methods, and details are omitted here for simplicity. For example, a voice feature parameter mel-frequency cepstral coefficient (MFCC) can be extracted from a voice signal of an audio stream, and then an eigenvector is calculated by using a method such as an i-vector (a speaker recognition algorithm) or a probabilistic linear discriminant analysis (PLDA, that is, a channel compensation algorithm for voiceprint recognition).

[0058] Step 404: Determine whether the voiceprint eigenvectors of the plurality of audio streams are consistent.

[0059] For example, when the user reads the user's ID at least twice during

registration, there are correspondingly at least two collected audio streams. To ensure that a difference between voiceprint features of the plurality of audio streams is not too big, voiceprint consistency determining can be performed between the plurality of audio streams. For example, a score of similarity between the plurality of audio streams can be calculated based on a voiceprint eigenvector extracted from each audio stream in step 403.

[0060] If the similarity score falls within a certain score threshold, it indicates that the audio streams satisfy a similarity requirement; and in this case, proceed to step 405. Otherwise, it indicates that there is a big difference between the plurality of audio entered by the user, and the user that is performing registration can be instructed to read the user's ID again, that is, to collect an audio stream again.

[0061] Step 405: Generate a model voiceprint feature based on the voiceprint eigenvectors of the plurality of audio streams.

[0062] In the present step, weighted summation can be performed on voiceprint eigenvectors extracted from audio streams in the previous step, to obtain the model voiceprint feature.

[0063] After the previous registration procedure is completed, the object registration information of the target object has been stored in the database. The object registration information can include an object identifier and a corresponding model physiological feature. The model physiological feature can include a model voiceprint feature and a model facial feature, and identity authentication processing of an object can be performed below based on the object registration information.

Identity authentication

[0064] FIG. 5 illustrates an example identity authentication procedure. In this procedure, a physiological feature used for authentication is described by using a combination of a facial feature and a voiceprint feature as an example. In addition, physiological features can be compared after it is determined that a target object that is being authenticated is a living object instead of a video. As shown in FIG. 5, the authentication procedure includes the following processing:

[0065] Step 501: Obtain a collected audio and video stream, where the audio and video stream is generated by a target object to be authenticated.

[0066] For example, a user can log in to an application with a relatively high

security requirement only after the user's identity has been authenticated by the application. In the present step, the user can start an application on an intelligent device, for example, a smartphone of the user, and the user can collect an audio and video stream to be authenticated by using a camera and a microphone of the smartphone. The audio and video stream can be generated when the user reads the user's application ID.

[0067] Step 502: Determine whether the lip reading and voice in the audio and video stream are consistent.

[0068] In this example, it can be first determined whether the lip reading and the voice in the audio and video stream are consistent. For a specific consistency determining procedure, references can be made to FIG. 2, and details are omitted here for simplicity.

[0069] If the lip reading and the voice are consistent, it indicates that the target object that is being authenticated is a living object instead of a video, etc. In this case, proceed to step 503. Otherwise, go back to step 501 to perform collection again.

[0070] Step 503: Perform voice recognition on an audio stream in the audio and video stream to obtain voice content of the audio stream. For example, the voice content obtained through recognition can be a user's ID "123456".

[0071] Step 504: Use the voice content as an object identifier of the target object, and determine whether pre-stored object registration information includes the object identifier.

[0072] For example, if the pre-stored object registration information includes the object identifier, a model physiological feature corresponding to the object identifier can be obtained in the object registration information, for example, a model facial feature and a model voiceprint feature. Physiological recognition is further performed on the audio and video stream to be authenticated, to obtain a physiological feature of the target object, so as to compare the physiological feature with the model physiological feature. If the pre-stored object registration information does not include the object identifier, the user can be notified that the user has not performed registration.

[0073] Step 505: Perform voiceprint recognition on the audio and video stream to obtain a voiceprint feature of the target object. For extraction of the voiceprint feature in the present step, references can be made to FIG. 4.

[0074] Step 506: Perform facial recognition on the audio and video stream to

obtain a facial feature of the target object.

[0075] Then the physiological feature of the target object can be compared with the model physiological feature to obtain a comparison result, and if the comparison result satisfies the authentication condition, it is determined that the target object has been authenticated. For example, step 507 to step 509 are included.

[0076] Step 507: Compare the voiceprint feature of the target object with a model voiceprint feature to obtain a voiceprint comparison score.

[0077] Step 508: Compare the facial feature of the target object with a model facial feature to obtain a facial comparison score.

[0078] Step 509: Determine whether the voiceprint comparison score and the face alignment score satisfy the authentication condition.

[0079] For example, it is determined that the target object has been authenticated if the voiceprint comparison score and the facial comparison score satisfy at least one of the following: the voiceprint comparison score is greater than a voiceprint score threshold, and the facial comparison score is greater than a facial score threshold; or a product of the voiceprint comparison score and the facial comparison score is greater than a corresponding product threshold; or a weighted sum of the voiceprint comparison score and the facial comparison score is greater than a corresponding weighted threshold.

[0080] If it is determined that the voiceprint comparison score and the facial comparison score satisfy the authentication condition in the present step, it is determined that the target object has been authenticated. Otherwise, it is determined that the target object fails to be authenticated.

[0081] In addition, in this identity authentication example, similar to the previous identity registration procedure, voice recognition can be performed on an audio stream to obtain a user's ID after it is determined that lip reading and voice are consistent, or a user's ID can be obtained when a syllable time point in an audio stream is identified. In the previous example, the user's ID is identified after it is determined that lip reading and voice are consistent.

[0082] According to the identity authentication method in this implementation of the present application, an audio and video stream needs to be generated only once during authentication of a user. For example, the user needs to read the user's ID only once. In the method, the user's ID can be obtained by performing voice recognition on the audio and video stream, and a facial feature and a voiceprint feature can also be

verified by using the same audio and video stream. It simplifies the user's operation, improves the authentication efficiency, maintains a one-to-one authentication model, and ensures recognition accuracy. In other words, the recognized physiological feature is compared only with a feature corresponding to an object identifier in a database, thereby ensuring recognition accuracy. In addition, in the method, consistency between lip reading and voice is determined to ensure that a target object is a living object instead of a fake video record of an attacker, thereby improving the authentication security and reliability. In the method, the user's ID and the physiological feature obtained through recognition are obtained based on the same audio and video stream. To some extent, a fake audio and video stream of an attacker can be recognized.

[0083] To implement the previous identity authentication method, an implementation of the present application further provides an identity authentication apparatus. As shown in FIG. 6, the apparatus can include an information acquisition module 61, an identifier determining module 62, an information management module 63, a feature recognition module 64, and an authentication processing module 65.

[0084] The information acquisition module 61 is configured to obtain a collected audio and video stream, where the audio and video stream is generated by a target object to be authenticated.

[0085] The identifier determining module 62 is configured to determine whether the lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, use voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object.

[0086] The information management module 63 is configured to obtain a model physiological feature corresponding to the object identifier from object registration information, if the pre-stored object registration information includes the object identifier.

[0087] The feature recognition module 64 is configured to perform physiological recognition on the audio and video stream to obtain a physiological feature of the target object.

[0088] The authentication processing module 65 is configured to compare the physiological feature of the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication

condition, determine that the target object has been authenticated.

[0089] Referring to FIG. 7, in an example, the feature recognition module 64 can include a voiceprint recognition submodule 641 and a facial recognition submodule 642.

[0090] The voiceprint recognition submodule 641 is configured to perform voiceprint recognition on the audio and video stream to obtain a voiceprint feature of the target object.

[0091] The facial recognition submodule 642 is configured to perform facial recognition on the audio and video stream to obtain a facial feature of the target object.

[0092] The authentication processing module 65 is configured to compare the voiceprint feature of the target object with the model voiceprint feature to obtain a voiceprint comparison score, and compare the facial feature of the target object with the model facial feature to obtain a facial comparison score; and if the voiceprint comparison score and the facial comparison score satisfy the authentication condition, determine that the target object has been authenticated.

[0093] In an example, it is determined that the target object has been authenticated if the voiceprint comparison score and the facial comparison score satisfy at least one of the following: the voiceprint comparison score is greater than a voiceprint score threshold, and the facial comparison score is greater than a facial score threshold; or a product of the voiceprint comparison score and the facial comparison score is greater than a corresponding product threshold; or a weighted sum of the voiceprint comparison score and the facial comparison score is greater than a corresponding weighted threshold.

[0094] In an example, as shown in FIG. 7, the identifier determining module 62 can include the following: a syllable recognition submodule 621, configured to identify a voice syllable and a corresponding time point in the audio stream in the audio and video stream, and identify a lip reading syllable and a corresponding time point in a video stream in the audio and video stream; and a consistency determining submodule 622, configured to determine that the lip reading and the voice are consistent if both the voice syllable and the lip reading syllable are consistent with the corresponding time point.

[0095] In an example, the information acquisition module 61 is further configured to obtain an audio and video stream to be registered of the target object.

[0096] The identifier determining module 62 is further configured to use voice content obtained by performing voice recognition on an audio stream in the audio and video stream as the object identifier of the target object, when the lip reading and the voice in the audio and video stream to be registered are consistent.

[0097] The feature recognition module 64 is further configured to perform physiological recognition on the audio and video stream to be registered, to obtain the model physiological feature of the audio and video stream to be registered.

[0098] The information management module 63 is further configured to correspondingly store the object identifier of the target object and the corresponding model physiological feature in the object registration information.

[0099] The previous descriptions are merely example implementations of the present application, and are not intended to limit the present application. Any modifications, equivalent replacements, improvements, etc. made within the spirit and principle of the present application shall fall within the protection scope of the present application.

CLAIMS

1. An identity authentication method, the method comprising:

(a) obtaining an audio and video stream, wherein the audio and video stream is generated by a target object to be authenticated;

(b) determining whether lip reading and voice in the audio and video stream are consistent;

(c) using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object;

(d) determining whether pre-stored object registration information comprises the object identifier;

(e) obtaining a model physiological feature corresponding to the object identifier from pre-stored object registration information, if the pre-stored object registration information comprises the object identifier,

wherein determining whether lip reading and voice in the audio and video stream are consistent, using voice content obtained by performing voice recognition as the object identifier of the target object, and obtaining the model physiological feature corresponding to the object identifier from object registration information are executed in parallel;

(f) performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object;

(g) comparing the physiological feature of the target object with the model physiological feature to obtain a comparison result;

(h) determining whether the comparison result satisfies an authentication condition; and

(i) in response to determining that the comparison result satisfies an authentication condition, determining that the target object has been authenticated.

2. The method according to claim 1, wherein the physiological feature comprises a voiceprint feature and a facial feature, and the model physiological feature comprises a model facial feature and a model voiceprint feature.

3. The method according to claim 2, wherein performing physiological recognition on the audio and video stream to obtain a physiological feature of the target object comprises:

performing voiceprint recognition on the audio and video stream to obtain a voiceprint feature of the target object; and

Date reçue / Date received 2021-12-09

performing facial recognition on the audio and video stream to obtain a facial feature of the target object.

4. The method according to claim 3, wherein comparing the physiological feature of the target object with the model physiological feature to obtain a comparison result, and if the comparison result satisfies an authentication condition, determining that the target object has been authenticated, comprises:

comparing the voiceprint feature of the target object with the model voiceprint feature to obtain a voiceprint comparison score;

comparing the facial feature of the target object with the model facial feature to obtain a facial comparison score;

determining whether the voiceprint comparison score and the facial comparison score satisfy the authentication condition; and

in response to determining that the voiceprint comparison score and the facial comparison score satisfy the authentication condition, determining that the target object has been authenticated.

5. The method according to any one of claims 2 to 4, wherein it is determined that the target object has been authenticated if the voiceprint comparison score and the facial comparison score satisfy at least one of the following:

the voiceprint comparison score is greater than a voiceprint score threshold, and the facial comparison score is greater than a facial score threshold; or

a product of the voiceprint comparison score and the facial comparison score is greater than a corresponding product threshold; or

a weighted sum of the voiceprint comparison score and the facial comparison score is greater than a corresponding weighted threshold.

6. The method according to claim 1, wherein determining whether lip reading and voice in the audio and video stream are consistent, and if the lip reading and the voice are consistent, using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as an object identifier of the target object comprises:

identifying a voice syllable and a corresponding time point in the audio stream in the audio and video stream;

identifying a lip reading syllable and a corresponding time point in a video stream in the audio and video stream; and

Date reçue / Date received 2021-12-09

determining that the lip reading and the voice are consistent if both the voice syllable and the lip reading syllable are consistent with the corresponding time point.

7. The method according to claim 1, wherein before obtaining a collected audio and video stream, the method further comprises:

obtaining an audio and video stream to be registered of the target object;

in response to determining that the lip reading and the voice in the audio and video stream to be registered are consistent, using voice content obtained by performing voice recognition on an audio stream in the audio and video stream as the object identifier of the target object;

performing physiological recognition on the audio and video stream to be registered to obtain the model physiological feature of the audio and video stream to be registered; and

storing the object identifier of the target object and the corresponding model physiological feature in the object registration information.

8. The method according to claim 1, wherein the target object to be authenticated comprises a user of a user device, the audio stream comprises a reading voice of the user, the video stream comprises a reading video image of the user, and a collected audio and video stream comprises an audio and video stream generated when the user registers with an application.

9. An identity authentication apparatus comprising:

an information acquisition module configured to perform step (a) of claim 1;

an identifier determining module configured to perform step (b) of claim 1;

a feature recognition module having a voiceprint recognition submodule and a facial recognition submodule, where:

the voiceprint recognition submodule is configured to perform step (c) of claim 1;

the facial recognition submodule is configured to perform steps (d), (e) and (f) of claim 1; and

an authentication processing module configured to perform steps (g), (h) and (i) of claim 1.

Date reçue / Date received 2021-12-09

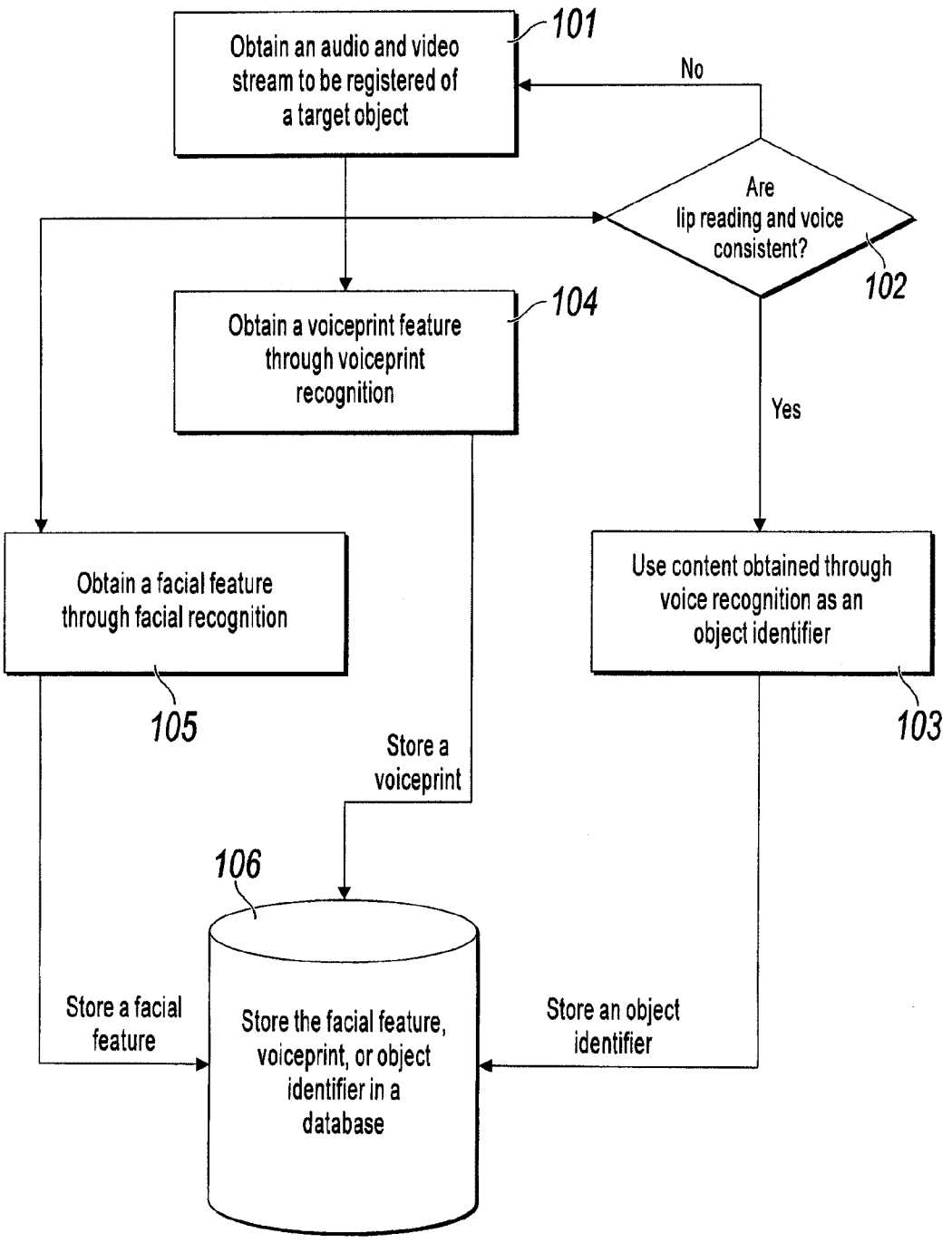

Obtain an audio and video stream to be registered of a target object
101
No
Are lip reading and voice consistent?
102
Yes
Obtain a voiceprint feature through voiceprint recognition
104
Obtain a facial feature through facial recognition
105
Use content obtained through voice recognition as an object identifier
103
Store a voiceprint
106
Store a facial feature
Store the facial feature, voiceprint, or object identifier in a database
Store an object identifier

FIG. 1

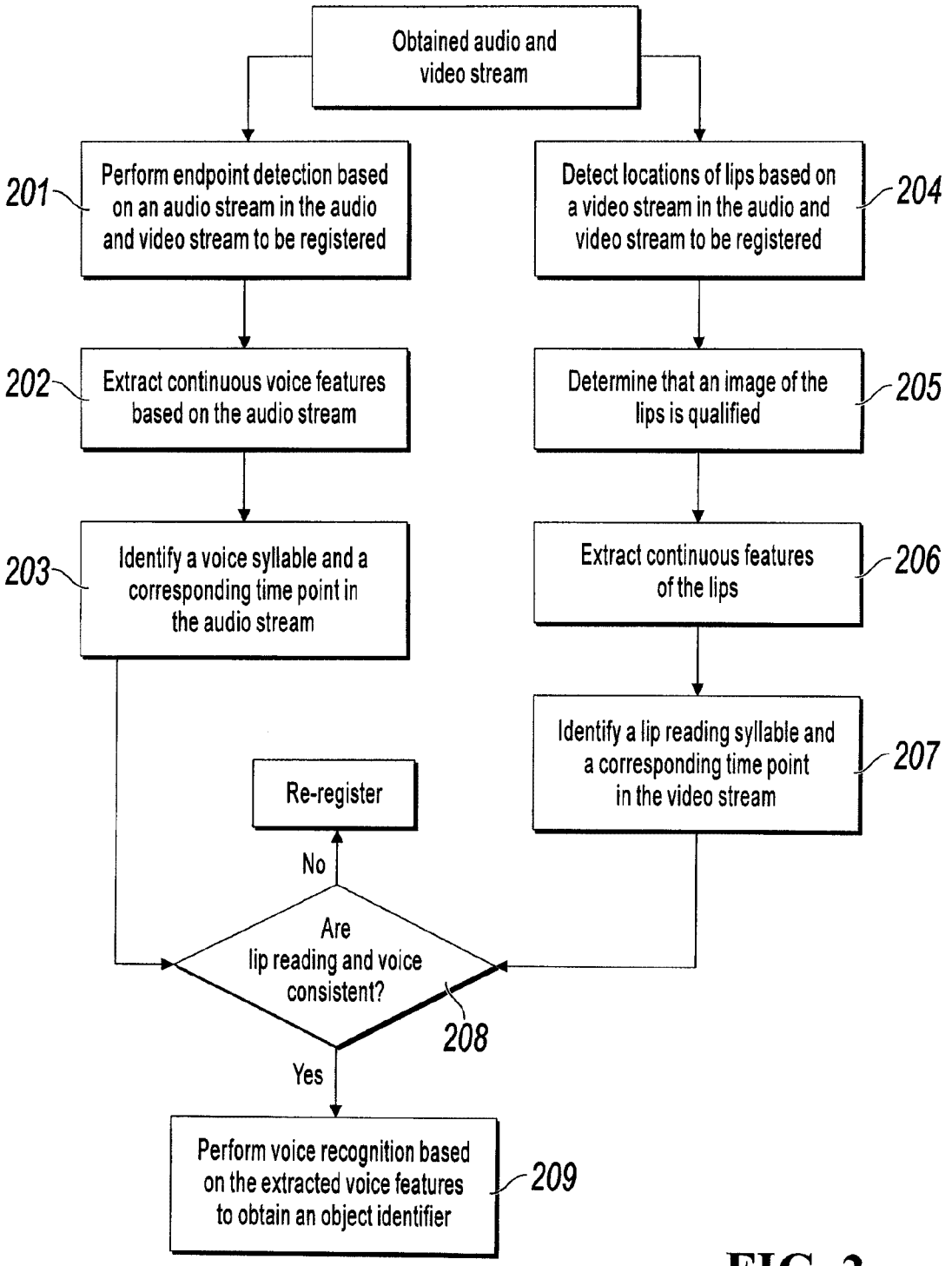

Obtained audio and video stream
Perform endpoint detection based on an audio stream in the audio and video stream to be registered
201
Extract continuous voice features based on the audio stream
202
Identify a voice syllable and a corresponding time point in the audio stream
203
Detect locations of lips based on a video stream in the audio and video stream to be registered
204
Determine that an image of the lips is qualified
205
Extract continuous features of the lips
206
Identify a lip reading syllable and a corresponding time point in the video stream
207
Re-register
No
Are lip reading and voice consistent?
208
Yes
Perform voice recognition based on the extracted voice features to obtain an object identifier
209

FIG. 2

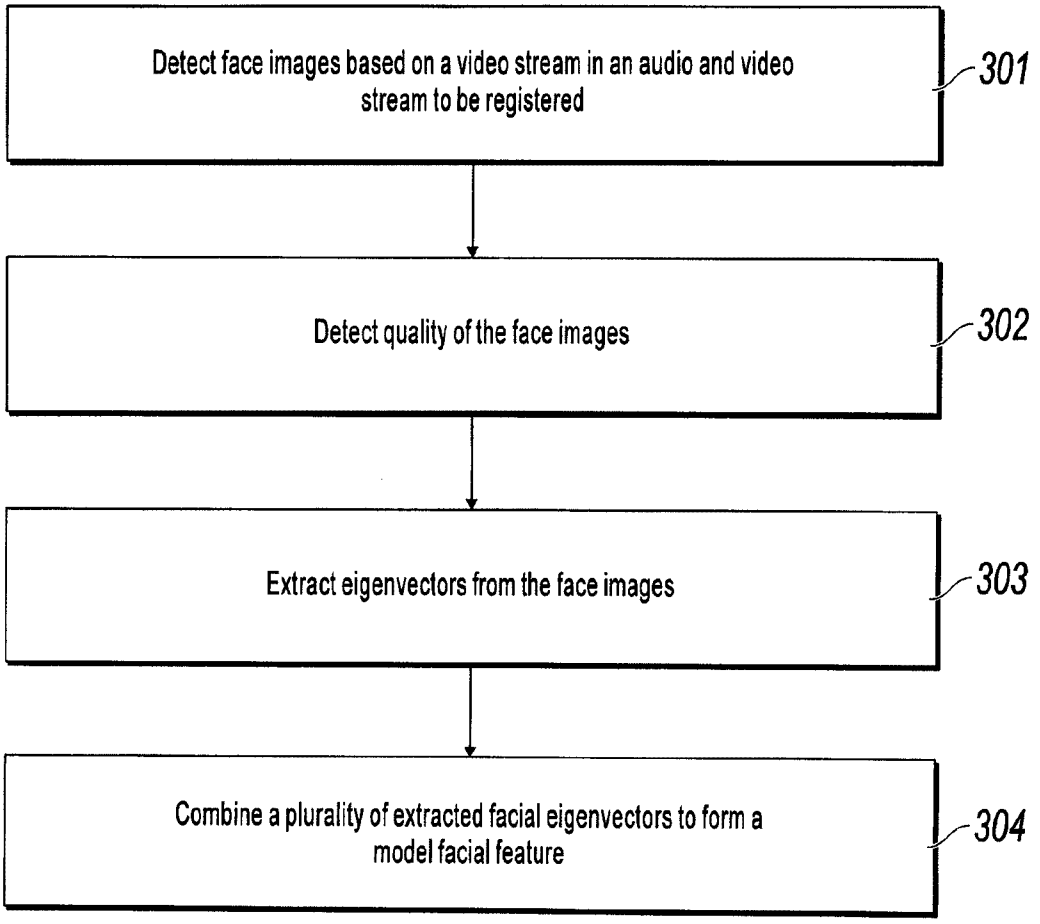

Detect face images based on a video stream in an audio and video stream to be registered
301
Detect quality of the face images
302
Extract eigenvectors from the face images
303
Combine a plurality of extracted facial eigenvectors to form a model facial feature
304

FIG. 3

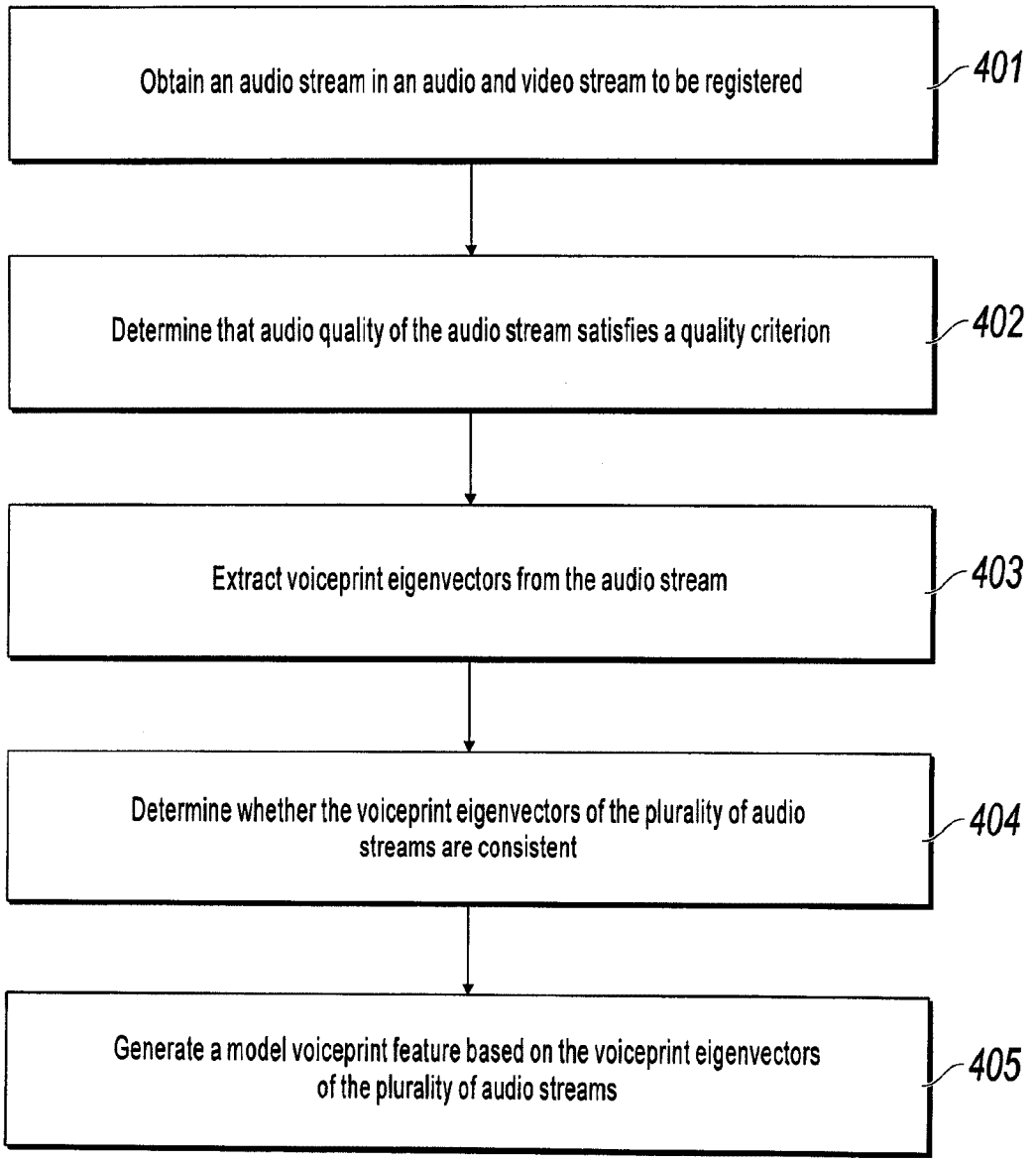

Obtain an audio stream in an audio and video stream to be registered
401
Determine that audio quality of the audio stream satisfies a quality criterion
402
Extract voiceprint eigenvectors from the audio stream
403
Determine whether the voiceprint eigenvectors of the plurality of audio streams are consistent
404
Generate a model voiceprint feature based on the voiceprint eigenvectors of the plurality of audio streams
405

FIG. 4

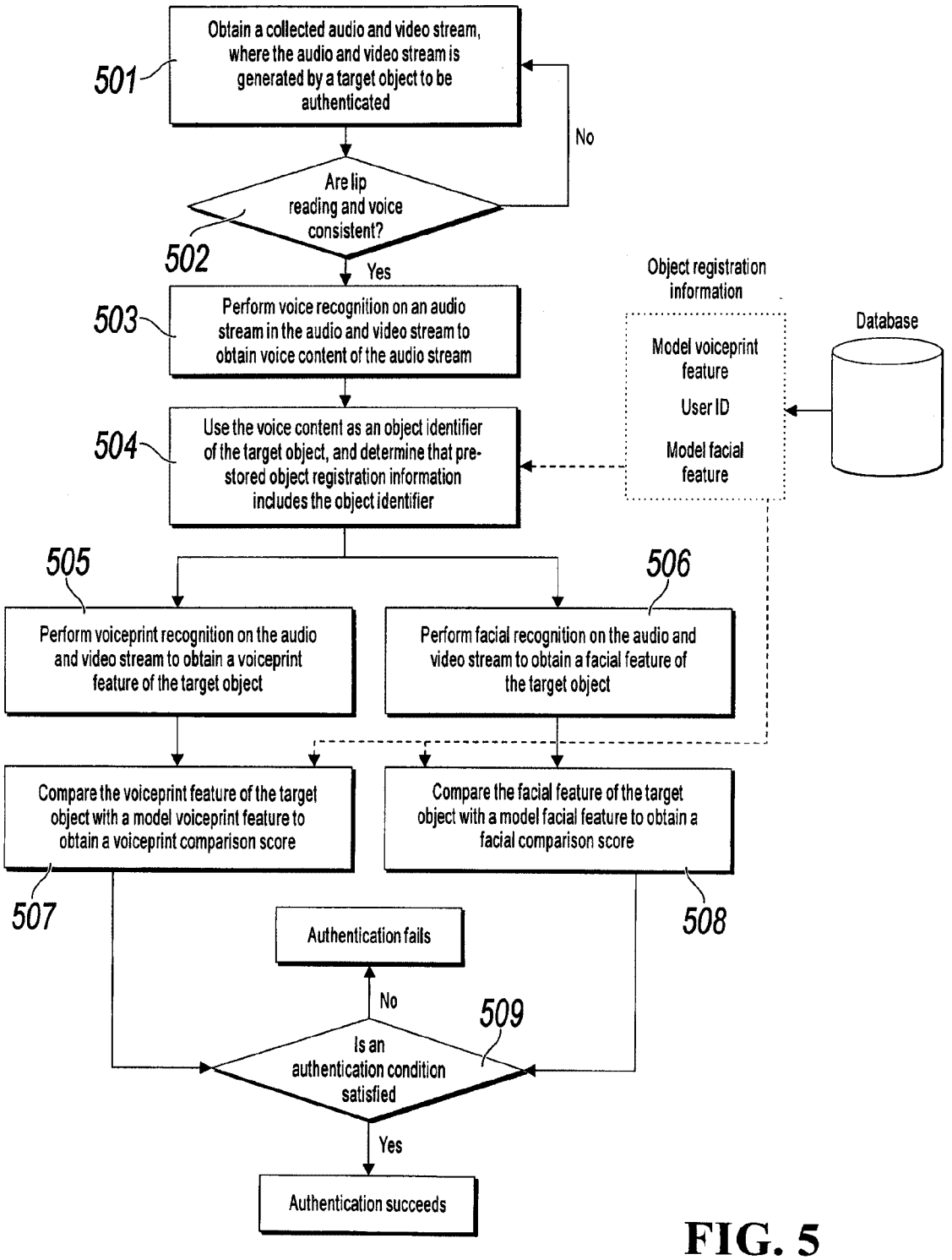

Obtain a collected audio and video stream, where the audio and video stream is generated by a target object to be authenticated
501
Are lip reading and voice consistent?
No
502
Yes
Perform voice recognition on an audio stream in the audio and video stream to obtain voice content of the audio stream
503
Use the voice content as an object identifier of the target object, and determine that pre-stored object registration information includes the object identifier
504
Object registration information
Model voiceprint feature
User ID
Model facial feature
Database
505
Perform voiceprint recognition on the audio and video stream to obtain a voiceprint feature of the target object
506
Perform facial recognition on the audio and video stream to obtain a facial feature of the target object
Compare the voiceprint feature of the target object with a model voiceprint feature to obtain a voiceprint comparison score
507
Compare the facial feature of the target object with a model facial feature to obtain a facial comparison score
508
Authentication fails
No
509
Is an authentication condition satisfied
Yes
Authentication succeeds

FIG. 5

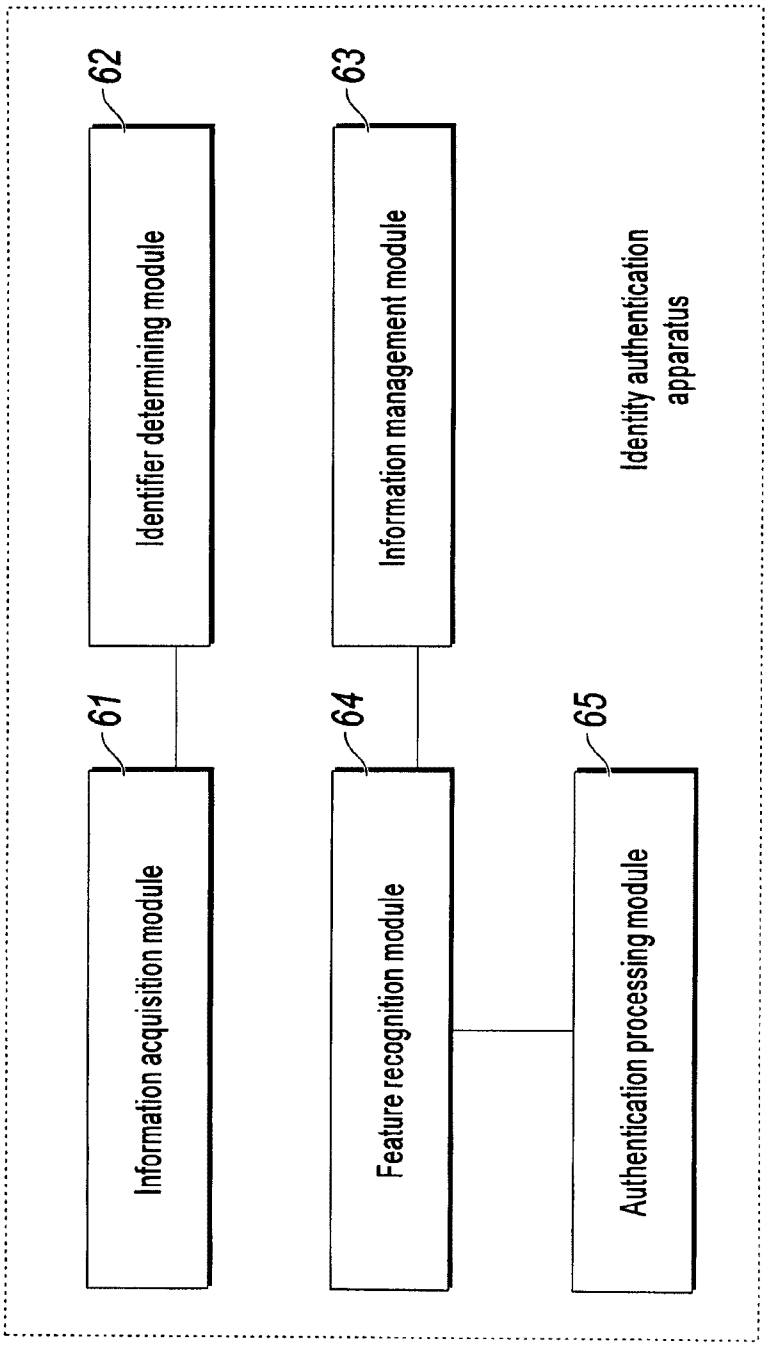

62
Identifier determining module
63
Information management module
Identity authentication apparatus
61
Information acquisition module
64
Feature recognition module
65
Authentication processing module

FIG. 6

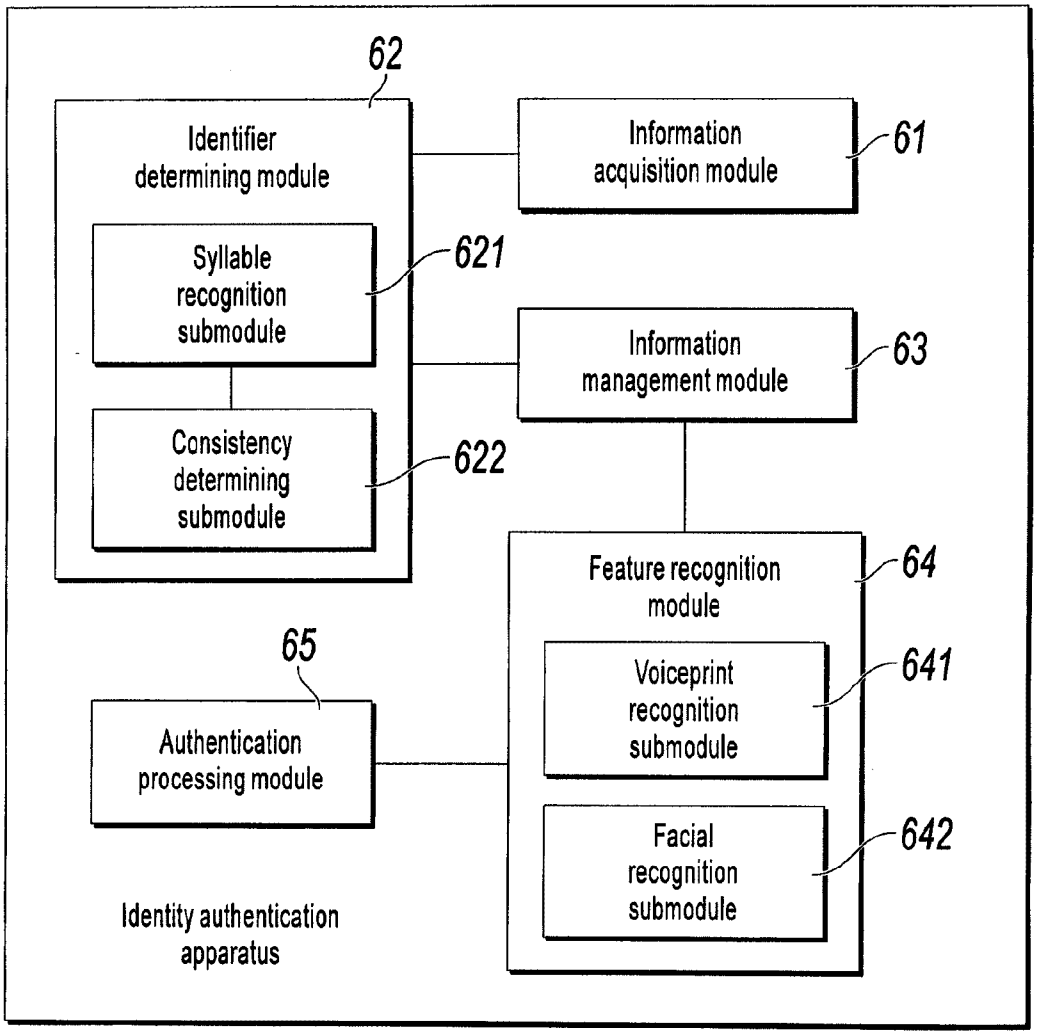
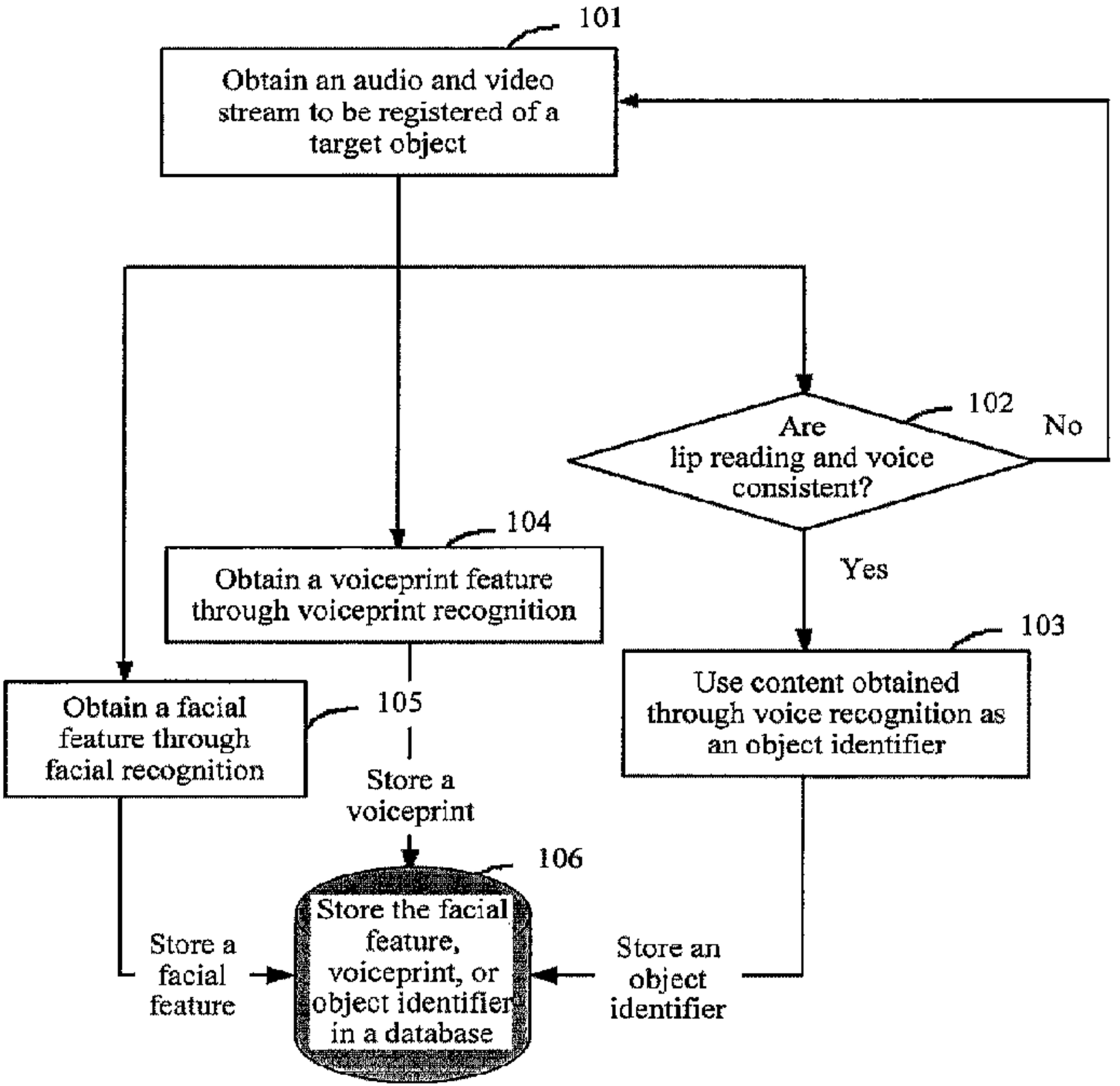

62
Identifier determining module
Syllable recognition submodule
621
Consistency determining submodule
622
Information acquisition module
61
Information management module
63
Feature recognition module
64
Voiceprint recognition submodule
641
Facial recognition submodule
642
65
Authentication processing module
Identity authentication apparatus

FIG. 7


101
Obtain an audio and video stream to be registered of a target object
102
Are lip reading and voice consistent?
No
Yes
104
Obtain a voiceprint feature through voiceprint recognition
103
Use content obtained through voice recognition as an object identifier
Obtain a facial feature through facial recognition
105
Store a voiceprint
106
Store the facial feature, voiceprint, or object identifier in a database
Store a facial feature
Store an object identifier