Identity Authentication based on Audio Visual Biometrics: A Survey

Kai Li
Department of Computer Science
University of Central Florida
Orlando, Florida 32816
Email: kaili@eecs.ucf.edu

Abstract—**Biometric authentication is an emerging technology that utilize biometric data for the purpose of person identification or recognition in security applications. A number of biometrics can be used in a person authentication system. Among the widely used biometrics, voice and face traits are most promising for pervasive application in every life, because they can be easily obtained using unobtrusive and user-friendly procedures. The ubiquitousness of low-cost audio and visual capture sensors on smart phones, laptops, and tablets has made the advantages of voice and face biometrics more outstanding compared with others. For quite a long time, the use of acoustic information alone has been a great success for speaker authentication applications. Meanwhile, the last decades or two also witnessed great advancement in face recognition technologies. However, in adverse operating environments, neither of these techniques achieves optimal performance. Since visual and audio information conveys correlated and complimentary information to each other, integration of them into one authentication system can potentially increase the system's performance, especially in suboptimal operating conditions. In this paper, I made an extensive survey on state-of-the-art authentication technologies based on the fusion of audio and visual biometrics. The major components of an audio-visual biometric system will be firstly discussed. Then various aspects of different existing biometric systems will be analysed and compared. Finally, a novel idea of dynamic 3D audio-visual biometric authentication exploiting the Microsoft Kinect device will be presented.**

I. Introduction

Biometrics can be defined as measurable characteristics of the individual based on his physiological features or behavioural patterns that can be used to recognize or verify his identity. A physiological biometric would identify a person by iris scan, DNA or fingerprint. Behavioural biometrics are related to the behaviour of a person, including but not limited to: typing rhythm, gait, and voice. Biometric identifiers are used in computer science as a form of automatic identity authentication and access control. With the emergence of smart phones and third and fourth generation mobile and communication devices, and the appearance of a ”first generation” type of mobile devices with biometric identity verification, such biometric authentication technology has received a great amount of attention for safety and security in all aspects of our daily lives, and many governments are heavily funding biometric research.

Fig. 1: Commonly used biometrics. Top row: voice, signature, fingerprint, hand geometry. Bottom row: face, iris, key stroke dynamics, DNA

Token-based identification systems, such as a driver's license or passport, and knowledge-based identification systems, such as a password or personal identification number have been used for a long time as the routine way of obtaining the permission for accessing security systems. However, those methods can be easily breached and are therefore unreliable to some degree, leaving the attackers a lot of room to initiate various attacks. Things will be much better if we consider biometric traits as our own passwords, because it would be much more difficult for the attackers to simulate or copy their targets' biometric cues to access restricted security systems. Moreover, biometric characteristics are intrinsic to their owners and therefore can hardly be borrowed, stolen or forgotten [4].

A number of biometric characteristics can be used in identity authentication systems, such as fingerprint, iris, DNA and face, gait, voice etc. Several issues need to be considered in designing and applying biometric systems: accuracy of identification, robustness to spoof or imposter attacks, user acceptance, and cost of capture sensors etc. Among those factors, user acceptance and sensor cost are the primary obstacles that prevents such highly accurate and robust biometrics as DNA, iris from entering civilian daily lives, because they are generally considered obtrusive and require sophisticated and expensive devices. In comparison, face and voice biometrics are the most user-friendly and cost-effective choices. As a matter of fact, we human beings determine the identity of a person primarily by his face or voice. Therefore, it's also natural to utilize them for person identification. However, face and voice based person identification are not panacea for all kinds of security applications. For example, the false acceptance rate of face and voice based biometric system is normally in the order of 10^{-2}, which is far from enough for applications of high security requirements. While for many other biometric applications, such as sport venue entrance check, PC access, building access, which do not require high security, face and voice provides ideal choice for unobtrusive and low-cost automatic identity authentication.

Face and voice recognition technologies are for a long time advancing independently. Great successes have been achieved on both research areas. For example, in a controlled environment with high signal-to-noise (SNR) levels, the speaker recognition rate is approximately hundred percent [1] (less than 1% error). And face recognition rate can consistently stay above 95% in well-controlled environment. However, both experience degraded performance in suboptimal operational conditions. Speaker recognition systems that uses only audio information are susceptible to microphone types, background and channel noises, and acoustic environments. While face recognition systems are rather sensitive to illumination conditions, background changes, speaker occlusions, image or video qualities etc. Another weakness common to both techniques is that they are vulnerable to imposter attacks. The attacker can play the recorded or synthesize target's voice at the time of voice-based authentication, or show an image or video of the target's face in front of a face recognition system. This vulnerability render face and voice based single-modality authentication system powerless in front of spoof attacks.

Multiple modal biometric systems is suggested by researchers to overcome the weakness of a single modality biometric system [5]-[15]. An audio-visual biometric authentication system utilize the speech together with static or sequences of video frames containing the face or part of the face (e.g. the mouth area), in order to improve the person recognition performance and increase the robustness of the system to spoof attacks. The combined and simultaneous use of audio and visual information provides a greater degree of security as tampering any one of these sources would not be enough for false access and authentication. In addition, since multiple modalities (e.g. audio modality and visual modality) provides complimentary information of the same audio-visual event (e.g. the utterance of words or sentences in the authentication process), unfavourable operating conditions for one modality might be compensated by the performance of the other. Another advantage of the audio-visual multi-modal biometric is that the the integration of the two brings in almost zero extra cost, because audio and visual sensors almost always come in pair, for example, in cellphones, laptops, digital cameras and so on.

TABLE I: A Comparison of different Biometrics

	Accuracy	**User Friendliness**	**Ease of Use**	**Cost**
DNA	High	Low	Low	High
Retina	High	Low	Low	High
Iris	High	Low	Low	High
Fingerprint	High	Medium	Medium	Medium
Voice	Medium	High	High	Low
Face	Low	High	High	Low
Signature	Medium	Medium	Low	High

In this paper, state-of-the-art audio-visual (AV) biometric systems will be reviewed. Various aspects of different AV biometric systems will be discussed. The contributions of this paper are threefold. Firstly, more than fifteen papers and technical documents are reviewed and summarized to give a systematic overview of the current research status of AV biometric authentication technologies. Secondly, the relative advantages and disadvantages of different types of AV biometric systems are compared and analysed. Thirdly, a novel idea that exploits the newly emerging low-cost devices such as Kinect for dynamic 3D AV biometric authentication is proposed and some initial study findings are presented.

The remainder of the paper will be organized as follows. Section II gives an overview of the major components of a general AV biometric system and briefly introduces the AV biometric identity authentication process. In Section III, various commonly used AV features will be introduced and analysed. Section IV-A discusses different kinds of AV fusion techniques. In Section V, several existing AV biometric systems are described and compared. Section VI presents the novel idea of utilizing Kinect device for dynamic 3D AV biometric authentication, as well as some initial study results. Finally, the paper will be concluded in section VII.

II. AV Biometric System Overview

A. AV Biometric System Classification

There are several ways to categorize AV biometric systems. In terms of whether users are required to say specific words or sentences during authentication, AV biometric systems can be classified into Text-Dependent (TD) and Text-Independent (TI) ones. Text-dependent AV biometric system require users to repeat all or a portion of a fixed amount of pre-enrolled words at authentication time. While text-independent systems recognize persons solely based on the acoustic characteristics of the user. Therefore, the user can say anything in the authentication process.

AV biometric systems can also be categorized with respect to the type of visual information they use. AV biometric systems that utilize static visual information (face images or static video frames containing faces) are called Audio-Visual-Static (AVS) biometric systems. While those utilizing visual features containing temporal information obtained from video sequences are called Audio-Visual-Dynamic (AVD) biometric systems.

B. AV Biometric System Components

Whichever category an AV biometric system belongs to, it contains similar major components as depicted in Fig. 2. The first component is the preprocessing of audio and visual signals. The preprocessing of audio signal involves speech denoising [10], silence detection[1] and removal [2] and signal enhancement. Visual signal processing consists of face detection, face tracking (in AVD systems), or specific facial region extraction (e.g. mouth region or the entire lower face). The second component is the feature extraction component. Feature extraction is a critical for the performance of AV biometric systems. Well chosen audio and visual features should have good discriminative power to differentiate between different individuals, as well as robustness to changing environment conditions. A lot of research has been done in extracting various audio and visual features, which will be covered in detail in Section III. It's worthy to note that the preprocessing and feature extraction component are not totally independent components. The preprocessing of audio and visual signals should refer to and serve the choice and extraction of the appropriate features.

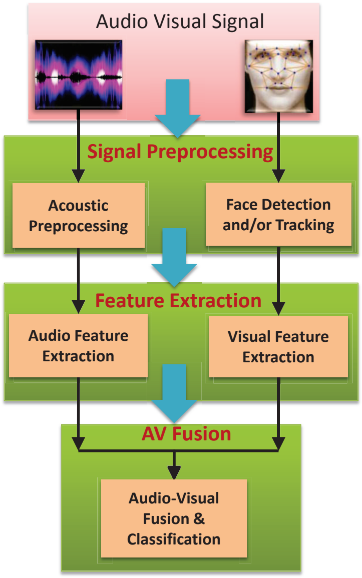

Fig. 2: The major components of an AV biometric system, adapted from [4]

AV fusion is a special case of multi-modal biometric research, which concerns the study of effectively integrating multiple biometric characteristics for higher person identification performance than each of its constituent single modality system. The increased system performance includes both person identification rate and robustness to environment changes. The fusion of audio and visual information can take place at various levels of a AV biometric system, and can be roughly classified into early fusion and late fusion. Early fusion techniques combine the audio and visual information at the feature level. Early fusion has to deal with the difference between video and audio sampling rates. Depending on the video standard, frame rates vary between 24 to

[1] The silence part of the signal largely affects the performance of the system. In fact, silence does not carry meaningful information of the speaker, while its existence causes biased score thus worsen the system performance.

60 frames per second. While audio sampling rates are typically in the range of tens of kilo hertz. The common method of handling the sampling disparity is to down-sample the audio signals or up-sample the video signal. Late fusion takes place at the score or decision level. Score fusion typically maps the scores from different modalities to a common interval and combines them by weighted summation or weighted product. The final score will be compared with a pre-defined threshold to reject or accept the person being authenticated. In decision fusion methods, each modality will firstly make a rejection or acceptance decision independently and then their decision results will be combined with majority voting or with AND/OR operation.

C. *Authentication Process*

There are two different phases of operation for an AV biometric system: (1) Enrolment and (2) Authentication. In the enrolment phase (see Fig. 3), audio and visual information from users will be added to the audio visual database. Typically, multiple samples of a users' audio visual information will be collected to train the classifiers. In template matching based classification, one or several prototypes computed from the collection of a user's audio visual samples will be stored in the database for authentication use. In the authentication phase, the live audio visual information from the users are captured and compared against the records stored in the database to yield the authentication results. In some systems, it's possible to automatically update the prototype template after each valid authentication, so that the system can adapt to gradual minor changes of a user's audio and visual characteristics, for example, as a result of ageing.

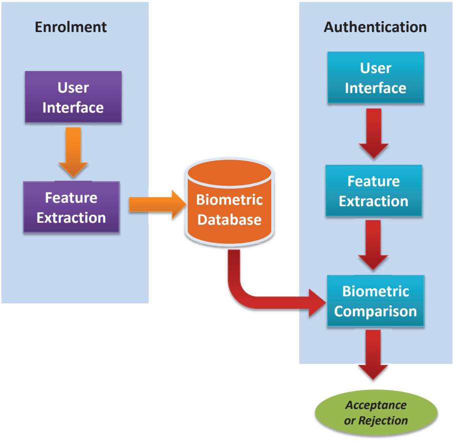

Fig. 3: AV biometric system operations

III. AV Feature Extraction

A. *Audio Features*

A number of acoustic features have been used in the literature for speaker recognition in both clean and noisy speech conditions. The most commonly used features are mel-frequency cepstral coefficients (MFCCs)[14] and linear prediction coefficients (LPCs)[3]. The processing steps for MFCC is shown in Fig. 4. The high-pass filter is used to enhance those high frequencies in the speech signals that are generally attenuated during the speech recording process. After applying frame blocking and Hamming windowing [2], short term speech segments with predefined length and overlap are processed by Fast Fourier Transform (FFT) and Mel-filter banks to get Mel spectrum. The Mel spectrum are transformed using Discrete Cosine Transform (DCT) to obtain Mel spectrum. Finally, vector quantization will be applied get compressed MFCCs.

LPC estimates current values of a discrete-time speech signal $x(n)$ by the linear combination of previous samples

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$$

where $\hat{x}(n)$ is the predicated signal value, $x(n-i)$ is the previous observed values and a_i are the LPC coefficients. The error generated by this estimation is

$$e(n) = |x(n) - \hat{x}(n)|$$

These LPC coefficients can then be solved by minimizing the estimation error.

Audio features are sometimes augmented by their first-order or second-order derivatives. The inclusion of derivatives captures the temporal dynamics of audio signals and provides more useful information. The proper choice of audio features depends on the operating conditions (e.g. background noise, acoustic environment)

[2] Investigations show that speech signal characteristics stays stationary in a sufficiently short period of time interval (It is called quasi-stationary). For this reason, speech signals are processed in short time intervals. It is divided into frames with sizes generally between 30 and 100 milliseconds. Each frame overlaps its previous frame by a predefined size. The goal of the overlapping scheme is to smooth the transition from frame to frame. The second step is to window all frames. This is done in order to eliminate discontinuities at the edges of the frames. If the windowing function is defined as $w(n)$, $0 < n < N-1$ where N is the number of samples in each frame, the resulting signal will be $y(n) = x(n)w(n)$. Generally hamming windows are used.

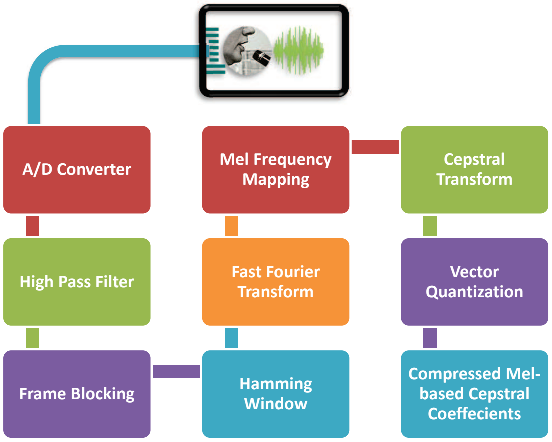

Fig. 4: Steps for computing MFCC audio features

and the fusion and classification algorithms used in later stages of the AV biometric system.

B. Visual Features

Compared with audio features, visual features used in AV biometric systems are much more abundant in quantity. The variety of visual features are generally grouped into three categories: (1) appearance-based features, (2) shape-based features, and (3) the combination of appearance and shape based features.

Appearance-based features directly consider the pixel values of the extracted face or mouth Region of Interest (ROI). The ROIs are typically small image patches containing mouth/lip region or the entire face extracted from a video frame. The rectangular region can also be extended to a 3D box containing a number of consecutive frames. The 3D representation captures the dynamics of visual speech information. Such acquired appearance features are typically in extremely high dimensions, which prohibits effective statistical modelling of visual features for later pattern classification. The side effect caused by high-dimensional feature vectors are usually termed as ”curse of dimensionality”. The overcome this problem, dimension reduction techniques are applied. Dimension reduction techniques, such as Principle Component Analysis (PCA) [8][9], Linear Discriminant Analysis (L-DA) [10][8] and Discrete Cosine Transform (DCT) [7], impose a linear transformation on the original feature vector, with the objective to produce a feature vector with a significantly lower dimension, while preserving most of the information carried by the original vector. PCA provides low-dimensional representation optimal in the mean-squared error sense, while LDA projects vectors into the most discriminative dimensions.

Shape-based features uses geometric or model-based (or template-based) representation of faces or lip contours. Geometric features are such features as height, width and area of the mouth. They are obtained by firstly extracting the mouth region and then locating the feature points. Model-based representations are generally more complex. A typical examples is snakes. A snake is an elastic curve connected by a set of control points. During training phase, the snake control points are iteratively updated to minimize some energy function. Template method is another way of representing the geometrics of faces. Another popular technique is Active Shape Model (ASM) [16]. ASM extracts a set of points from the ROI as the feature vector. The feature space are then transformed using PCA to get the axes of large shape variation. Active Appearance Model (AAM) [17] is an extension of ASM that capture the appearance variation of the region around the desired shape.

Both type of visual features have their relative merits and weaknesses. Appearance-based features contain the low-level information about the face and mouth movements. The extraction of appearance features are more simple and straightforward compared with shape-based methods. While the disadvantages of appearance-based visual features are that they are typically of much higher dimension than shape feature and that they are susceptible to illumination changes. In comparison, shape-based features capture the high-level information of face or mouth, and they are more robust to lighting variation because they focus on edge and contour information of facial parts. However, the computationally efficient and robust extraction of shape features remains a challenging problem in AV biometric research. In view of their relative strengths and weaknesses, the combination of both types of visual features are also considered by some researchers [18]. All in all, the choice of appropriate visual features for an AV biometric system should be based on a variety of factors such as operating environment, computational requirement, video quality etc.

IV. AV Fusion and Classification

In this section, I introduce different kinds of audio visual fusion techniques, as well as the classification methods commonly used in an AV biometric authentication system.

A. *Audio Visual Fusion*

AV fusion is a special case of multi-modal information fusion, where the fused modality are from audio and visual channels. Audio and visual signals provide unilateral description of the same audio-visual event. The fusion of audio and visual information been proved to improve the biometric system classification performance under various operating conditions [6]. AV fusion can be classified into early fusion and late fusion elaborated as follows.

- **Early Fusion.** In early fusion, audio and visual information are combined before classification. Specifically, audio and visual features are extracted independently from audio and visual signals. Then different features are combined by weighted summation or concatenation. Typically, audio and video streams are synchronized, and up-sampling of video or down-sampling of audio are performed before their features are combined.
- **Late Fusion.** Late fusion can be further classified to score-level fusion and decision-level fusion. Score-level fusion aims at combining the confidence scores of the models constructed from different features, in which each confidence score measures the possibility of classifying a test sample into positive class by a specific model. Weighted summation or weighted product can be used to compute the fused score. In decision-level fusion techniques, each modality will first independently classify the test sample as positive or negative. The final decisions are obtained by using AND or OR logical operators. For example, in AND fusion, a test sample will be accepted only if both audio and visual classifiers give positive decision. In OR fusion, a test sample will be accepted if either classifier gives positive decision.

In general, integrating information at an early stage is more effective than at later stage 4, because the features extracted from different biometrics can present much more information than those in other fusion levels.

B. *Audio Visual Classification*

AV Biometrics based authentication is essentially a classification problem. In a single user system, the authentication of the client is a two-class classification problem, namely the valid user class and the impostor class. In a multi-user biometric system, the number of classes is equal to the number of users (in closed-set biometric system) or one more than the number of users (in an open-set biometric system). In open-set system, the additional class represent the unknown class or impostor class. During authentication, the live audio visual features from the client will be compared against each of the classes. Various techniques are used for the classification problem, including Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVMs) and Probabilistic Neural Network (PNN). For example, in GMM-based classifiers, the posterior possibility of the client belonging to each of the classes are computed, and the class with the largest posterior probability will be chosen to give the final decision. If the chosen class is the impostor class, the client will be rejected. Otherwise, he will be accepted. SVM and PNN capture the complex non-linear relationship between input features and class label. They are firstly be trained on a large corpus of positive and negative samples. Upon test time, the system will directly give a classification decision for the input test sample.

V. Comparison of different AV Biometric Systems

In this section, a number of existing AV biometric authentication system will be briefly introduced. In Table II, those systems are listed in time-ascending order and they are compared in terms of text dependency, audio & visual features, and fusion methods.

Ben-Yacoub et al. [5] developed both text-dependent and text-independent AV biometric system using audio information and frontal face information. Elastic graph matching is used to get face matching scores. Several binary classifiers including SVM, Bayesian classifier, Fisher's linear discriminant, decision tree and multilayer perceptron, are investigated for post-classification opinion fusion. The best results are obtained with SVM and Bayesian classifiers.

Aleksic et al. [6] develop a text-dependent AV biometric system with MFCC and its first and second order derivatives as the audio feature, and a 3 dimensional PCA projection of outer-lip contour-based shape feature as the visual feature. Similar to their handling of audio features, the visual features are also augmented with first and second order derivatives to capture dynamics of visual

Author	**Year**	**Visual Info. Type**	**Text Dependency**	**Features**		**Fusion Method**
				Audio	**Visual**	
Ben-Yacoub [5]	1999	static	TD & TI	LPCs	appearance-based	late Fusion
Aleksic [6]	2003	dynamic	TD	MFCCs	shape-based	early Fusion
Fox [7]	2003	dynamic	TD	MFCCs	appearance-based	early and late fusion
Nefian [8]	2003	dynamic	TD	MFCCs	appearance-based	late fusion
Sanderson [9]	2004	static	TI	MFCCs	appearance-based	early and late fusion
Micheloni [10]	2009	static	TD	MFCCs	appearance-based	late fusion
Chetty [11]	2010	dynamic	TD	MFCCs	shape-based	early and late fusion
Zheng [12]	2010	static	TI	MFCCs	appearance-based	early fusion
Asadpour [13]	2011	dynamic	TD	MFCCs	shape-based	early fusion
Yu [14]	2012	dynamic	TI	MFCCs	appearance-based	early fusion
Zhao [15]	2012	static	TI	MFCCs	appearance-based	early fusion

TABLE II: A Comparison of different AV Biometric Systems

features. They use a single-stream HMM to integrate dynamic audio visual features and performance classification experiments under various SNR levels (from 0 to 30 dB).

Nefian et al. [8] developed a text dependent AV biometric system. They model the temporal sequence of audio visual observations obtained from speech and the shape of the mouth using a set of coupled hidden Markov model (CHMM). The likelihood score obtained using CHMM is combined with the face recognition likelihood obtained using an embedded hidden markov model. They show in the experiments that the AV system improves the accuracy of the audio-only and video-only approaches at all levels of SNR ratio from 5 to 30 dB.

Chetty [11] developed an AV biometric person authentication system with liveness verification. The use MFCCs and lip-region eigenlip features as acoustic and visual features respectively. They perform fusion of audio and visual features at both feature level and score level. Two kinds of replay attack scenarios are tested in their experiments, namely ”static” replay attacks and the ”dynamic” replay attacks. The ”static” replay attack uses synthetic fake recordings and still images of the victim, while ”dynamic” replay attack uses an photo-realistic audio-driven facial animation as the visual input. Their results indicate that the robustness of an AV authentication system to static relay attacks are quite satisfactory with the best achievable Equal Error Rate (EER) of 0.31%. However, for sophisticated dynamic replay attacks, the best EER is 10.06%, which calls for more robust audio-visual-based authentication techniques.

Asadpour [13] proposed a model-based feature extraction method which employs physiological characteristics of facial muscles producing lip movements. This approach exploits the intrinsic properties of muscles such as viscosity, elasticity and mass which are extracted from the dynamic lip model. Features such extracted reduce the odds of imitation to the largest extent. A multi-stream pseudo-synchronized HMM training method is adopted to combine audio and visual features. The features are then applied to a Hidden Markov Model (HMM) AV identification system. The proposed approach is compared to other feature extraction methods including Kalman filtering, neural networks, adaptive network fuzzy inference system and Auto Recursive Moving Average (ARMA) and achieved superior performance.

Yu [14] proposed an audio visual based text-independent person recognition system that utilize still images and text-independent audio signals. MFCC is used as the audio features. The visual features are extracted using Pyramidal Gabor-Eigenface (PGE) algorithm. A framework of Probabilistic Neural Network (PNN) is developed to achieve feature fusion. The recognition rate achieved by the proposed approach achieves a better recognition rate than any of the single modality.

VI. AV Authentication with Depth Data

Most of the AV biometric system reviewed so far have a major weakness that they did not take fraudulent replay attack scenarios into consideration. Therefore, they are vulnerable to spoofing attacks by pre-recorded AV data samples of the target. This section first discuss the vulnerabilities of various types of AV biometric systems, and then propose to use dynamic 3D AV modelling technique to increase the robustness of AV biometric authentication system.

As discussed in previous sections, in terms of audio content of client's utterance at authentication time, AV biometric authentication systems are classified into text-dependent and text-independent ones. While in terms of the utilized visual information, they can be classified as visual-dynamic or visual-static ones. The combination of different classifications result in four kinds of different AV biometric systems, namely, text-dependent visual-static, text-dependent visual dynamic, text-independent visual-static, and text-independent visual dynamic. Table III give a brief comparison of those systems listed in ascending order of robustness to imposter attacks. Text-dependent visual-static systems can be most easily compromised by playing audio recordings and placing a still image of the victim in front of the system. The other types can also be compromised by attacks of different level of sophistication. In general, text-independent AV biometric systems are most robust to impostor attacks, because the uttered words at authentication is arbitrary and the authentication is totally based on acoustic features of the client. Moreover, its combination with visual information makes it more robust even in face of synthesized audio signals. However, even for text-independent visual-dynamic systems, the attacker can still employ the advanced techniques as mentioned by [11], which create artificial speaking character utilizing efficient photo-realistic audio-driven facial animations technique with near-perfect lip-syncing of the audio and several image key-frames of the speaking face video sequence.

TABLE III: A Comparison of different types of AV Biometric system with respect to robustness to replay attacks

System Type	**Example**	**Anti-attack Capability**	**Attack Scenarios**
TD VS	[5][10]	Low	Replay audio recordings and present still images of the target
TD VD	[6][7]	Medium	Replay video recordings of the target
TI VS	[9][12]	Medium	Play Synthesized voice of the target and present his photo
TI VD	[14]	High	Use photo-realistic audio-driven facial animations with perfect lip-syncing

An AV biometric authentication system is essentially a security measure that should be robust against various attacks. However, the research on improving such kind of robustness is extremely limited. Some researchers suggest using liveness checks [11] to counter against impostor attacks. However, the performance of those techniques in sophisticated attack scenarios deteriorate much compared with simple replay attacks, rendering them not mature enough for an real-world security system. Much more research in liveness check approaches has to be done.

In view of the limitation of current anti-attack techniques in AV biometric authentication system, I propose to use 3D dynamic visual information based audio-visual speaker model to perform speaker authentication. An intrinsic advantage of 3D face modelling approaches in audio-visual recognition task is that it's extremely difficult, if impossible, for the attacker to launch replay attacks, because those recordings do not carry any 3D information and will be rejected in the very beginning of the authentication process. To the best of my knowledge, there has been no published research work that utilizes 3D dynamic audio visual information for biometric authentication. I would like to briefly explain the idea and present some initial-study findings. The detailed design of the whole algorithm and the implementation and results is beyond the scope of this course-project. I will continue on current work and develop a paper out of it following the end of the class.

The Microsoft Kinect [19] provides a perfect platform for the 3D dynamic audio-visual person recognition. Kinect devices have been extremely popular since their appearance two years ago, due to their low-cost and availability. Kinect can capture VGA (640×480) resolution color image stream up to 30 frames per second, as well as a depth image stream with the same resolution and frame rate. The depth image pixel value are raw readings that represent the relative distance from the object to the Kinect sensor, which is proportional to the physical distance. An important first-step to make sure it could

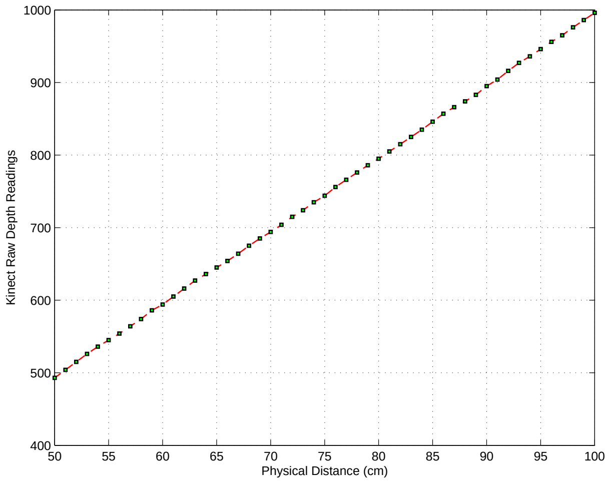

Fig. 5: The relationship between Kinect raw depth readings and the physical distance

be used for the proposed 3D audio-visual recognition task is to get the relationship between raw depth to the physical distance, because the face modelling require a precision at least in the order of centimetre. Fig. 5 shows this relationship in the range of 50 centimetres to 1 metre, which is a desired operational range for a typical biometric authentication system. The results are obtained by placing a flat board in the center of field of view of the Kinect and enlarge the horizontal distance by 1 centimetre every step and read the center value of the depth image. As can be seen from the plot, the the linearity in the 0.5 to 1m range is quite smooth, which is desirable for simple linear transformation between raw depth data and the physical distance for modelling. The raw depth readings vary by 1 when the physical distance changes by 1mm, which guaranteed enough resolution for detailed modelling of the entire face or the mouth and lips for audio-visual analysis.

VII. Conclusion

In this paper, I made and extensive survey of state-of-the-art of audio-visual biometrics based authentication system. Based the generalization and comparison of different existing systems, I firstly presented a brief overview of the major components of an general audio-visual biometric system and explained the authentication process. Then the commonly-used audio and visual features in AV biometric system and introduced, followed by the discussion of audio-visual fusion and classification. A tabular comparison of the existing systems are presented afterwards, with brief explanation for some of them.

I also propose a novel idea that exploits the Microsoft Kinect devices for 3D dynamic audio-visual person recognition, following analysing the comparative robustness of different audio-visual biometric systems. Some initial study findings are presented and explained. However, the full details of the proposed approach and the implementation are beyond the scope of this course project and will be carried out after the class with the goal to develop an academic paper.

References

[1] Campbell, J.P., Jr., ”Speaker recognition: a tutorial,” Proceedings of the IEEE , vol.85, no.9, pp.1437,1462, Sep 1997

[2] Greige, Hanna, and Walid Karam. ”Audio-Visual Biometrics and Forgery.”

[3] Atal, B. S. ”Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.” the Journal of the Acoustical Society of America 55 (1974): 1304.

[4] Aleksic, Petar S., and Aggelos K. Katsaggelos. ”Audio-visual biometrics.” Proceedings of the IEEE 94.11 (2006): 2025-2044.

[5] Ben-Yacoub, Souheil, Yousri Abdeljaoued, and Eddy Mayoraz. ”Fusion of face and speech data for person identity verification.” Neural Networks, IEEE Transactions on 10.5 (1999): 1065-1074.

[6] An Audio-Visual Person Identification and Verification System Using FAPs as Visual Features In Proceedings of the 2003 Workshop on Multimodal User Authentication (2003) by P. S. Aleksic, A. K. Katsaggelos

[7] Fox, Niall, and Richard B. Reilly. ”Audio-visual speaker identification based on the use of dynamic audio and visual features.” Audio-and Video-Based Biometric Person Authentication. Springer Berlin Heidelberg, 2003.

[8] Nefian, Ara V., et al. ”A Bayesian approach to audio-visual speaker identification.” Audio-and Video-Based Biometric Person Authentication. Springer Berlin Heidelberg, 2003.

[9] Sanderson, Conrad, and Kuldip K. Paliwal. ”Identity verification using speech and face information.” Digital Signal Processing 14.5 (2004): 449-480.

[10] Micheloni, Christian, Sergio Canazza, and Gian Luca Foresti. ”Audiovideo biometric recognition for non-collaborative access granting.” Journal of Visual Languages & Computing 20.6 (2009): 353-367.

[11] Chetty, Girija. ”Robust Audio Visual Biometric Person Authentication with Liveness Verification.” Intelligent Multimedia Analysis for Security Applications 10.1007 (2010): 59-78

[12] Haomian Zheng; Meng Wang; Zhu Li, ”Audio-visual speaker identification with multi-view distance metric learning,” Image Processing (ICIP), 2010 17th IEEE International Conference on , vol., no., pp.4561,4564, 26-29 Sept. 2010

[13] Vahid Asadpour, Mohammad Mehdi Homayounpour, Farzad Towhidkhah, Audiovisual speaker identification using dynamic facial movements and utterance phonetic content, Applied Soft Computing, Volume 11, Issue 2, March 2011, Pages 2083-2093.

[14] Yu, Chenxi, and Lin Huang. ”Biometric recognition by using audio and visual feature fusion.” System Science and Engineering (ICSSE), 2012 International Conference on. IEEE, 2012.

[15] Xuran Zhao; Evans, N.; Dugelay, J., ”Multi-view semi-supervised discriminant analysis: A new approach to audio-visual person recognition,” Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European , vol., no., pp.31,35, 27-31 Aug. 2012.

[16] Cootes, Timothy F., et al. ”Active shape models-their training and application.” Computer vision and image understanding 61.1 (1995): 38-59.

[17] Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor. ”Active appearance models.” Computer VisionECCV98. Springer Berlin Heidelberg, 1998. 484-498.

[18] Dupont, Stphane, and Juergen Luettin. ”Audio-visual speech modeling for continuous speech recognition.” Multimedia, IEEE Transactions on 2.3 (2000): 141-151.

[19] Microsoft Corp. Redmond WA. Kinect for Xbox 360. 1, 2