Detection of Face Spoofing Using Visual Dynamics

Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony T.S. Ho.

***Abstract*—Rendering a face recognition system robust is vital in order to safeguard it against spoof attacks carried out by using printed pictures of a victim (also known as print attack) or a replayed video of the person (replay attack). A key property in distinguishing a live, valid access from printed media or replayed videos is by exploiting the information dynamics of the video content, such as blinking eyes, moving lips, and facial dynamics. We advance the state of the art in facial anti-spoofing by applying a recently developed algorithm called Dynamic Mode Decomposition (DMD) as a general-purpose, entirely data-driven approach to capture the above liveness cues. We propose a classification pipeline consisting of DMD, Local Binary Patterns (LBP), and Support Vector Machines (SVM) with a histogram intersection kernel. A unique property of DMD is its ability to conveniently represent the temporal information of the entire video as a single image with the same dimensions as those images contained in the video. The pipeline of DMD+LBP+SVM proves to be efficient, convenient to use, and effective. In fact only the spatial configuration for LBP needs to be tuned. The effectiveness of the methodology was demonstrated using three publicly available databases: print-attack, replay-attack, and CASIA-FASD, attaining comparable results with the state of the art, following the respective published experimental protocols.**

***Index Terms*—DMD, spoofing, replay-attack, print-attack, CASIA-FASD, LBP, SVM.**

I. Introduction

A. Scope of study

The usage of biometrics in the area of security and for authentication purposes is of well-established importance. However, the widespread use of biometrics is hampered by security concerns such as database tampering [1] and sensor tampering (or presentation attack) [2], and many other attacks described by Ratha *et al.* [3]. While database tampering can lead to authorisation of an illegal individual, or denial of service to a legitimate user, sensor tampering can cause the granting of access right to an unauthorised user; thus, severely compromising the security and usability of biometric systems. In this paper, we are concerned with rendering a face recognition system robust against presentation attacks; and specifically, print attacks based on printed facial images, and replay attack that is carried out by replaying a video sequence using a tablet or a mobile phone.

The key innovation introduced here is the exploitation of the facial dynamic information in a completely data-driven fashion, rather than using prior knowledge regarding live face images such as eye-blinking and lip-movements (e.g. [4]). Since attack types are often unknown and very different from each other. Knowledge-driven approaches are unlikely to be able to scale well to a plethora of attacks. Each new attack potentially requires the algorithm designer to develop a specific cue that is peculiar to the attack being employed. Data-driven approaches, on the hand, can automatically learn to discriminate valid (i.e. live) videos[1] from attack videos given a sufficiently wide range of training data and attack types.

In order to extract facial dynamic information reliably, we propose a modified version of Dynamic Mode Decomposition (DMD). Since DMD has been used extensively in modelling fluid dynamics, we postulate that DMD can also capture the complex dynamics of head movements, eye-blinking, and lip-movements found in a typical video sequence containing face images. However, DMD has not been used for classification tasks before. For this reason, we propose a DMD-based classification pipeline involving a texture-based descriptor, coupled with a discriminative classifier. Our experiments show that DMD can indeed capture unique features that clearly distinguish videos produced by print and replay attacks from live (valid) videos containing an authentic face.

B. Background on attacks directed at biometric systems

Attacks directed at biometric systems can be broadly classified as database tampering and sensor attacks:

Sensor tampering which is also referred to as a presentation attack [6] is targeted towards fooling a biometric sensor. A common type of presentation attack is spoofing, which refers to the use of fake biometric data in order to deceive a biometric system [2]. Spoofing attempts dealing with fake fingerprints dates back to 1998, when Wills and Lees [7] indicated that four out of six devices they tested were vulnerable to fake prints attacks. Research has also been carried out on the use of gum-based fingerprints to fool biometric systems by Matsumoto *et al.* [8]. Thalheim *et al.* [9] further demonstrated that fingerprints, iris scans and face recognition systems are vulnerable to replay attacks.

For face recognition, there are two types of presentation attacks, namely print attacks and replay attacks. A print attack refers to facial spoofing carried out by presenting a printed photo to a camera [10]. Indeed, this attack is very easy to carry out because getting hold of a target's photo is extremely simple. With the increasing use of social media, few users consider the repercussions associated with an act as simple as uploading their pictures [10]. Replay attacks, on the other hand, are carried out by replaying a previously recorded face image (video) of a target user in order to spoof a biometric system. The video can be replayed easily using a hand-held mobile tablet. Again, with the widespread use of high-quality

S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki and A.T.S. Ho are with the Faculty of Engineering and Physical Sciences (FEPS), University of Surrey, UK. e-mail: ({s.tirunagari, n.poh, d.windridge, a.iorliam, n.suki and a.ho}@surrey.ac.uk)

Manuscript received July 7, 2014;

[1] We refer to live, authentic videos as *valid* videos following closely Pinto's terminology [5]; as opposed to *attack* (or *spoofed*) videos produced by various spoofing methods.

consumer-grade camcorders and social media networks, this type of attack can be carried out easily in both remote and logical access control systems protected by a face recognition system.

We exclude from this study replay attacks that require an attacker to know about the architecture of a biometric system. These attacks assume that the attacker can intercept the communication channel between the biometric sensor and its feature extractor, thus acquiring a raw biometric image; or intercept the communication between the feature extractor and the biometric matcher, thus acquiring an extracted biometric feature set. It is further assumed that the attacker can replay or reintroduce the copied data (raw image or biometric feature) into the respective communication channel in order to gain access to such a secured system [11]. These attacks are much more sophisticated to carry out and are best counteracted by securing the communication channels; therefore, they are out of the scope of this study.

C. Motivation for DMD

There is a considerable challenge in extracting relevant subcomponents or modes from video data. Gaining a deep and accurate understanding of complex flow behaviour requires a form of *reduced order models* [12] or *mode decomposition* [13]. The mode decomposition of the form that we are interested is commonly used in the area of computational fluid dynamics (CFD) [14]. Specifically, Dynamic mode decomposition (DMD), is a mathematical method developed to extract the relevant modes from empirical data generated by nonlinear complex fluid flows. These modes provide a reduced-order representation of the complex flow behaviour, called coherent flow structures. The extraction of these coherent flow structures is of great importance in CFD. In the context of face counter-spoofing, we conjecture that DMD has the potential of discovering the facial dynamics that captures the vitality signs of valid face videos whilst extracting artefacts of spoof videos such as moiring and planar effects, at the same time. The moiring patterns are often regarded as undesired artefact of images, the effect of which is similar to fibers in moire silk. The planar effect is the appearance of an picture that is flat rather than 3D, as opposed to a real face.

DMD works analogously to Principle Component Analysis (PCA). Whereas DMD contains dynamic information about the data under consideration, PCA lacks this property [15]. Tirunagari *et al.* [16] compared PCA and DMD in extracting Kelvin-Helmoltz instabilities from complex flows. They showed that DMD is superior to PCA in analysing complex fluid flows. This is consistent with the literature in both experimental and numerical flow field data problems [15]. Indeed, not only can DMD extract dynamic information effectively, thus, temporal dynamic characteristics, it can also capture spatially-coherent structures [15]. Schmidt *et al.*. [17] used DMD on a sequence of fluid flow images and illustrated how it can detect dynamically relevant coherent structures for characterising fluid behaviour over a given time interval. Schmid [15] introduced a DMD-variant by taking into consideration an approximation of the linear mapping for temporal 'snapshots' and afterwards detects the relevant frequencies. This method is, however, dependent on the snapshots acquired. Problems addressed by DMD include: determining the flow over a square cavity, calculating the wake structure behind a flexible membrane, and determining the instabilities observed in experiments of a jet passing between two cylinders [15].

Similarly, we conjecture that extracting relevant modes from videos in face recognition can provide a good separability between valid videos and the spoofed ones. Particularly, for biometric applications such as replay and print attacks, in which observations are carried out in a continuous pattern i.e. several consecutive images in a video sequence rather than a single snapshot of image extracted from a video, we propose a modified DMD method that is computationally efficient for handling large size videos. This constitutes the originality of our proposal in tackling both photo and video based spoof attacks.

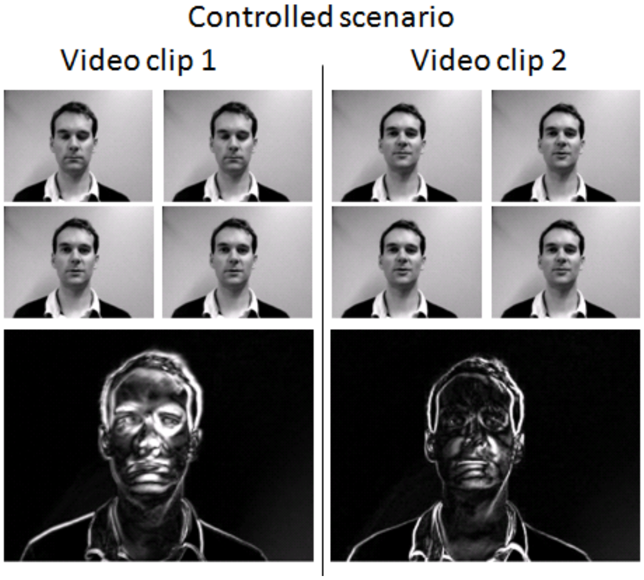

Fig. 1. Four frames from a valid access, controlled scenario video clips showing the eye-blinking and the lip movemnets. The bottom row shows their corresponding first dynamic mode image.

We demonstrate the effectiveness of DMD in Figure 1. The images in the top two rows show the corresponding extracted frames from each of the two videos containing a valid-access request in a controlled scenario. Since the person in these videos was speaking and blinking his eyes inevitably from time to time, one can observe that the corresponding first DMD mode highlights dynamic changes in the lip movements and eyes blinks. However, eye blinking is less obvious for the second video. It is, therefore, necessary to experimentally validate the effectiveness of DMD for this problem.

D. Contributions

Although much work has been directed towards tackling issues related to print attacks and replay attacks, as well as the development of publicly available databases, e.g., the print-attack database [18], the replay-attack database [19] and the CASIA-FASD database [20], there is still significant room for improvement for anti-spoofing methods in face recognition. In particular, there is no literature on the use of DMD on print and replay attacks. Our first contribution is, therefore, to introduce the use of DMD for countering spoof attacks in face recognition. Since DMD has not been used with classification of video sequences, we propose a system pipeline consisting of DMD, LBP and SVM. This combination is ideal because DMD captures the visual dynamics in the form of a fixed-size image, LBP can effectively capture the dynamic patterns, and SVM is known to be an ideal general-purpose classifier that minimizes the empirical risk of classification error. Extensive experiments were conducted on: (i) a print-attack dataset [18] containing 200 videos of real-access and 200 videos of spoof attempts using printed photographs of 50 different identities. (ii) Replay-attack dataset containing 200 videos of valid-access and 1000 videos of spoof attempts using printed photographs, tablets replays and mobile phones replays of 50 different identities. (iii) The CASIA-FASD containing 50 real clients with 150 valid-access videos and 450 attack videos.

Our second contribution is related to making the DMD more practical. First we implemented a version of DMD that is faster. While the original DMD uses QR-decomposition [21] (decomposition of a matrix in to an orthogonal matrix Q and an upper triangular matrix R) and singular value decomposition (SVD) methods [22], we propose to use LU decomposition [23] (factors a matrix as the product of a lower triangular matrix L and an upper triangular matrix U). Second, in our modified DMD, we use the absolute value of the complex modes when rendering the “DMD image”. Third, we recommend that the optimal mode to select the DMD modes using the zero phase angle criteria. In short, we have made DMD practical for classification and have demonstrated its effectiveness on separating valid access videos from spoofed ones.

Our final contribution is to simplify the taxonomy of existing facial anti-spoofing methods by categorising them as being data-driven or cue-based. For each category, a number of methods are further enumerated (see Figure 2) and supported by the relevant literature as shown in Table I. In this way, the strengths of different solutions proposed in the literature can be taken full advantage in order to improve the liveness detection task of the 2D face recognition technology.

E. Organisation

Section II, presents state-of-the-art methods using our proposed categorisation. Section III, presents the modified version of Dynamic Mode Decomposition (DMD) method and its application on spoof detection. A brief overview of the selected face spoofing datasets is presented in Section IV. Section V presents the experimental procedure including the results, followed by conclusions and future directions in Section VI.

II. Existing solutions

According to Chakka *et al.* [24], techniques to counter spoofing in 2-D face recognition systems are generally divided into motion, liveness, and texture analyses. Motion generally refers to motion features such as optical flow, e.g., [4] and [25]. Liveness analysis refers to the detection of the vitality signs of a biometric sample, such as head and lips movement, and eye-blinking. Finally, texture analysis refers to the use of texture descriptors for discriminating valid videos from the spoofed ones.

We argue that such a categorisation is ambiguous and does not help one to identify the strength of each technique. For example, a liveness analysis method that relies on movements is often implemented using motion descriptors such as optical flow. As a result, it is futile to argue whether or not a method is classified as belonging to liveness analysis or motion analysis.

Instead, we advocate to “decompose” existing methods into primary components in such a way that the strengths of different solutions proposed in the literature can be taken full advantage for improving the liveness detection task of the 2D face recognition technology. Our proposed broad categorisation of methods classifies a method as either data-driven or cue-based. This categorisation and some of the potential subcategories and specific methods are shown in Figure 2. While cue-based methods rely on intuitions and/or observations, data-driven methods tend to use generic image-processing algorithms that are applicable to a wide range of computer vision tasks. It is worth noting that the cue-based and data-driven attributes are not mutually exclusive. Indeed, a cue-based method may rely on generic texture descriptors such as local binary pattern (LBP) or motion descriptors such as optical flow. As a result, categorising a method as being texture-only or motion-only would not help a deeper understanding of which component of the method works, why it works, what assumption is made, and when it may fail. By decomposing a method based on cue-based attributes and

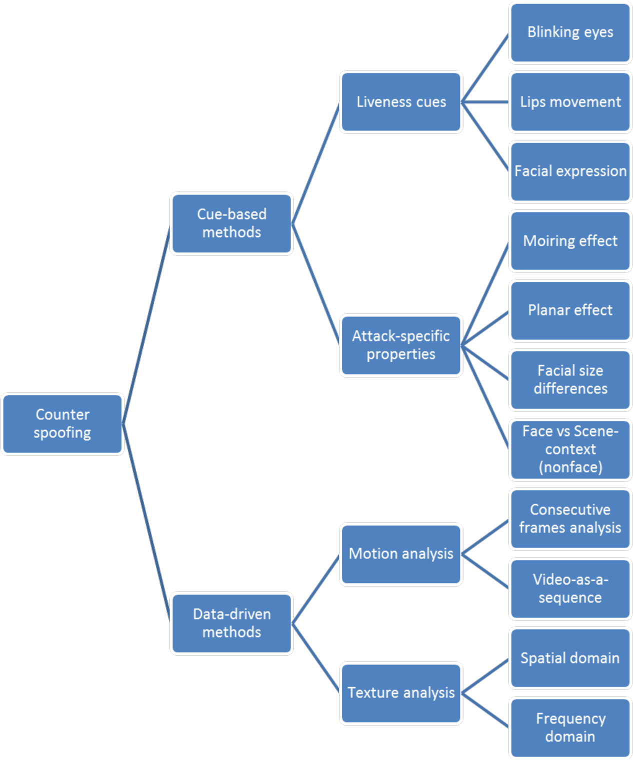

Fig. 2. Categorisation of specific methodologies existing in the literature.

data-driven attributes, the success or failure of a technique can be clearly identified. The result of this analysis is shown in Table I.

A. Cue-based methods

We further distinguish two types of cues, namely liveness cues and attack-specific properties/cues.

Liveness cues. We define liveness cues as ones that depend on the vitality signs of a biometric trait. For 2D face recognition, these cues are eye blinking, lips movement, and changes in facial expression. These cues were either made implicit or explicit by authors such as Tronci *et al.* [30], Li *et al.* [1], Tan *et al.* [27], Peixoto *et al.* [28], Pan *et al.* [33], and potentially many other authors. While this approach may be effective for countering attacks based on printed face images, it may fail completely with a video-based replay attack because a video can capture all aspects of facial liveness stated above. Given the ease of acquiring facial video and the high accessibility of facial videos in the public domain such as Facebook, often posted by the user himself, the possibility of an attacker employing a video-based replay attack is extremely high. In order to address this weakness, one will need to look for additional attack-specific cues that are described below.

Attack-specific cues are those that do not rely on the facial properties and are often idiosyncratic to attack method. An example of this is the moiring effect which is a secondary and visually undesirable artefacts that are caused by two superimposed patterns. In the context of 2D face attack, this effect is caused by the recapture of an image from another display device. This effect was pointed out in Pinto *et al.*'s work [5]. Interestingly, the non-face information such as the moiring effect can be obtained from the facial region but this is still considered an attack-specific cue. Another cue that is peculiar to a planar media such as a printed face image or a face image shown on a digital tablet is the flatness of the replicated image. The first use of this cue is attributed to Bao *et al.* [32].

Generality of cue-based methods. Because cue-based methods exploit the peculiarity of a specific attack setting or method, the solution may not always generalized to other attack methods. As a result, one may obtain poorer results than those reported in the initial studies. Indeed, some literature [5] suggests that cue-based methods that aim to exploit printed medium attacks may not work on replay video-based attacks or 3D mask attacks. For instance, Li's method [1] which is based on facial size difference and invariant facial pose and expression may not work when the assumptions do not hold. One way to invalidate the assumptions thus defeating the countermeasure is by rotating or bending a printed face photo with high very resolution, as suggested by Pinto *et al.* [5]. Another example is Pan's method [33] which is based on the cue that the face region and the region outside the face (that is, the scene context), are different. Since the background of an image may be different from one live video to another, the method may fail. Furthermore, with a 3D mask, an attacker can also simulate eye blinking by looking through a perforated 3D mask. Indeed, in an attempt to replicate Bao's method [32] and Kollreider's method [25], Anjos *et al.* [35] found that these methods did not work well for video-based replay methods even after tuning the parameters associated with the algorithms as optimally as possible on their database following a strictly well-defined and unbiased experimental protocol.

B. Data-driven methods

Data-driven methods can be further divided into motion and texture analyses. By motion analyses, we understand that the method exploits consecutive images in a sequence such as optical flow, or considering a video as a multidimensional sequence or time-series. In contrast, a method is considered texture analysis if it can derive information from a single image frame in the video.

For motion analysis, one can further divide them into two subcategories, using one of the following strategies:

1) **Consecutive-frame analysis**: analysing motion on two consecutive video frames, such as methods based on optical flow.

2) **Video-as-an-image-sequence analysis**: analysing motion by treating a video as a multidimensional sequence. Under this categorisation, Bao's method [32], Kollreider's method [25] and Anjos's method [35] that are based on optical flow are considered consecutive-frames analysis methods. Similarly, our method based on DMD also belongs to this category. However, DMD does so by taking into account of all possible consecutive pairs of images in a video sequence in its formulation (to be discussed in Section III-A). In this sense, video-as-an-image-sequence analysis is a generalisation of consecutive-frame analysis.

Video-as-an-image-set analysis. Another possible case that was not covered above is analysing a video by treating it as a collection or set of images. Such an approach typically does not consider the motion information. An example of this approach is to project a set of images into a texture subspace using principal component analysis (PCA). Since the ordering of the image does not matter, the motion information is not exploited.

Unlike PCA, DMD treats a video as a sequence of images and projects them in the principal motion subspaces. As a result, DMD is superior to PCA in classifying motion. Indeed, a large body of literature in fluid dynamics have demonstrated the superiority of DMD over PCA. For this reason, we conjecture that DMD is a viable candidate for classifying facial video liveness.

Video-as-a-set texture analysis. An extension of texture analysis is to apply a method to the entire set of video images. This leads to a variant of texture analysis that we refer to as video-as-a-set texture analysis. This type of analysis can be done at the feature-level or at the score level. In the first case, new features are derived by combining the features extracted from each image in a video sequence or a batch or chunk of video sequence. The derived features can be combined until an entire video is processed. LBP-TOP is one such example [36]. At the score level, typically the output of a liveness classifier which reflects a hypothesis score is used in the combation process. The classifier first classifies each image in a video

TABLE I
CATEGORISATION OF EXISTING FACIAL COUNTER-SPOOFING METHODS.

Broad category	**Data-driven**			**Cue-based (motivated by)**	
Narrow category	**Motion analysis**		**Texture analysis**	**Liveness cues**	**Attack-specific proper-ties**
	Consecutive-frame analysis	*Video analysis*	*Frequency or spatial analysis*	*E.g. blinking eyes, lips movement*	*E.g. planar effect, moiring effect*
Li's method [26]			2D Fourier spectrum	Photo attack has a smaller face size; and the facial expression and pose are invariant	
Tan's method [27]			Variational retinex-based and difference-of-Gaussian (DoG)	The surface roughness in terms of Lambertian reflectance between valid and spoof videos are different	
Peixoto's method [28]			Applying an adaptive histogram equalisation to the images before extracting latent reflectance features	The brightness of the LCD screen affects the recaptured image, which makes the image edges more susceptible to blurring	
Maatta's method [29]			Micro-texture analysis via LBP with SVM as a classifier		
Tronci's static method [30]			Texture histogram capturing colour in RGB and HSV space; edge directivity; MPEG descriptors	The loss of information through the image recapturing process and the peculiar noise	
Tronci's video-based method [30]				eye blinks, mouth alterations and changes in facial expression	
Schwartz's method [31]			shape, colour and texture of the face		
Bao's method [32]	Optical flow		Comparing a half of face with another half that is divided in two ways: horizontally and vertically.	Planar object movement cues aiming to detect four basic rigid object motion types, namely, translation, in-plane rotation, panning or out-of-plane rotation (swing)	
Kollreider's method [4] and [25]	Optical flow			Capturing the subtle movements of different facial parts (i.e., face and ear proportion), assuming that facial parts on real faces move differently than on photos.	
Pan's method [33] and [34]	Optical flow			Inside face changes (e.g., eye blinking) versus outside face (the scene context)	
Pinto's visual rhythm [5]		Multiple video spectra represented by gray-level co-occurrence matrices represented by 12 statistical descriptors	Frequency analysis of noise residuals		Noise residuals capture moiring effect
Anjos's method [35]	Optical flow-derived statistics based on Chi-square distance of a pair of Von-mises distributions			Inside face versus outside face context	
Our proposed method	Matrix decomposition of consecutive pairs of detected face images	Dynamic mode decomposition	LBP features classified by SVM	Although not based on cues, DMD can detect eye movements, lips movement, and the facial dynamics.	

sequence and the resultant scores, one from each image, are combined via a fixed operator such as mean or a second-level classifier that combines the statistics (e.g. [37]).

C. Key considerations in 2D facial counter-spoofing

The growing body of literature in 2D facial counter-spoofing have implicit addressed two key issues, namely:

- The importance of motion information.
- The need for a classifier to discriminate between valid and spoof facial videos.

Motion-based methods versus non-motion-based methods. Through a number of comparative studies, such as those reported by Anjos *et al.* [35], Chakka *et al.* [24], Chingovska *et al.*'s works [19], there is sufficient evidence to suggest that motion information is vital and methods based on this information can cover a wide range of attacks targeting the 2D face recognition technology. Indeed, motion can capture not only liveness cues but also attack-specific cues as discussed above. Later we will compare both types of approaches in our experiments, i.e., PCA, which represents a nonmotion-based method versus DMD, which represents a motion-based method.

Trainable versus non-trainable methods. Another important consideration among all the methods surveyed here is whether or not a method is trainable. All data-driven methods are considered trainable by definition because each of these methods requires a classifier to be able to discriminate valid videos from the spoofed ones. Although these methods can potentially be more general and more robust, they can only work well with spoofing methods that are found in the training set. If an attack type is not present in the training set, it is possible that the classifier may not generalise well. In comparison, cue-based methods may not require a classifier; therefore, they are considered non-trainable. An example of this is Anjos's optical flow correlation algorithm (in [35]) which does not need a trainable classifier but nevertheless requires a few parameters to be tuned via cross-validation on a given data set.

In summary, although the generalisation ability of a method to an unseen attack type is not guaranteed, trainable methods have the additional advantage that given a sufficiently representative training data that contains a plethora of attack methods, they are a preferred solution because they can exploit the additional data without significantly modifying the underlying algorithm unlike the cue-based methods. For this reason we consider a trainable methodology based on DMD.

D. Characteristics of DMD and its comparison with existing methods

1) Characteristics of DMD: Among the methods surveyed, few methods are completely data-driven and do not rely on any cues. For example, Pinto's method [5] is considered a cue-based method and data-driven because it exploits the moiring effect, the derived features of which are then classified using SVM. DMD is a method that qualifies both criteria. Indeed, DMD is an algorithm that exploits the motion information in a video sequence. Although optical flow can also capture this information for a pair of consecutive frames, DMD does so at the sequence level.

Although completely data-driven, we find that DMD can capture liveness cues and attack-specific artefacts at the same time. For instance, DMD can capture eye-blinking and lips movements, as well as the peculiarity of attacks. Furthermore, DMD can also capture the motion due to attack-specific artefacts such as planar movement. This is shown in Figure 6.

2) Competing methods: As explained, DMD is somewhat unique in its category of algorithms because of its ability to process multidimensional temporal sequence, which is a generalisation of consecutive-frames analysis such as algorithms based optical flow. Among them, we have chosen Anjos's method [35] which is considered superior among all the optical flow methods. We have also chosen Pinto's method [5] because of its use of a classifier. Table II compares the DMD with these two algorithms in terms of a number of attributes.

Pinto's visual rhythm works well only if the assumption about the high frequency image component, that is, the noise artefacts introduced by the recaptured printed/planar medium behave in the same way. Unfortunately, the video replay attack carried out by Pinto was based on LCD screen for replaying the video with a mounted (Sony Cybershot) camera. Therefore, no random movement that is typically produced by hand-held attack media is observed in the test set. It is, therefore, not clear how this movement might affect Pinto's algorithm.

Similarly, Anjos's optical flow correlation works well under the assumption that face and non-face regions have different distribution properties in terms of optical flow. Presumably a talking face video sequence would induce large chi-square distances. Anjos's method also implicitly relies on involuntary eye-blinking motions which happen every 3-4 seconds. However, a valid, face video without movement that happens between two eye-blinking motions would induce much smaller chi-square distances just as a printed facial photo attack does, leading to potential false rejection. A strength of their method is that they can use a standard face detector without explicitly needing to localise the two eyes in an image.

In comparison, the proposed DMD is completely data-driven; yet it is able to detect liveness cues and attack-specific artefacts, as described above. However, DMD needs a classifier. For this reason, we propose to use DMD in conjunction with LBP as a texture descriptor of the visual dynamics, and SVM as a back-end classification engine. This proposed DMD-pipeline was tested on both printed-medium attacks and replayed video attacks across various settings, including hand-held and mounted reply media, using the latest print-attack and replay-attack face datasets. Our experimental results, following strictly the published experimental protocols, show that the proposed DMD pipeline achieved the best performance reported so far (with the exception of komulainen's method [38] on CASIA-FASD dataset, to be further discussed in Section V-H) in comparison to the literature, i.e., Chingovska *et al.*'s work [19]. Indeed, the classification performance was perfect for the printed media and for the video-replay attacks, we achieved an HTER of 0.55% on the development set and 0% on the test set.

TABLE II
COMPARISON OF THE DMD WITH PINTO'S VISUAL RHYTHM AND ANJOS' METHOD IN TERMS OF A NUMBER OF ATTRIBUTES.

Attributes	**DMD pipeline (our proposal)**	**Pinto's visual rhythm [5]**	**Anjos' method [35]**
Consider multiple frames?	Video-as-an-image-sequence	Video-as-a-set	Video-as-an-image-sequence (see text)
Observational feature	Spatial domain (pixel)	Frequency domain of noise residual	Optical flow
Does observation capture motion?	Yes, for the entire sequence	Yes, via optical flow and their subsequent concatenated images (visual rhythm)	Yes, via optical flow
Features used for classification	LBP applied to DMD principal modes	12 statistical features derived from the gray-level co-occurrence matrices of the concatenated image-set spectra	A pair of Von-mises distributions derived from the facial region and non-face region of the optical flow output
Classifier	SVM	SVM, Partial Least Square	Chi-square distance
Cues relied upon or extracted	Automatically detect liveness cues and attack-specific artefacts	Authors' intuition on the moiring effect of the replay video	Authors' intuition that face versus nonface optical flow are different

III. METHODOLOGY

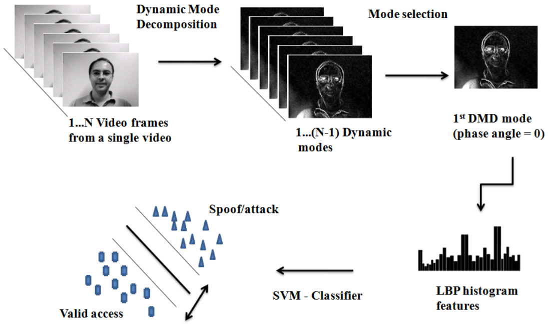

Fig. 3. Flow chart showing the steps involved in the methodological pipeline.

In this section, we present the pipeline of our method which consists of DMD, Local Binary Pattern histograms and a kernel based Support Vector Machine (SVM) classifier. The overall methodology or process pipeline is shown in Figure 3. First, a video is processed using the DMD algorithm in order to output dynamic mode images. From which, we select a single dynamic mode image corresponding to eigenvalue whose phase angle is = 0 or closest to it. Second, LBP histogram features are computed for this dynamic mode image. Finally, the produced LBP code is fed into a trained SVM classifier in order to classify whether the processed video is a valid access or spoof. Half Total Error Rate (HTER) is used to evaluate the performance measure. To validate our DMD pipeline we have used principle component analysis (PCA) based on snapshot approach as a baseline method.

A. Dynamic Mode Decomposition (DMD)

Let p_r be the r^{th} video frame whose size is $m \times n$. This video frame p_r is converted to $mn \times 1$ column vector, resulting in the construction of a data matrix **P** of size $mn \times N$ for N video frames.

$$\mathbf{P} = [p_1, p_2, p_3, \cdots, p_N] = \begin{pmatrix} p_1^1 & p_2^1 & \dots & p_N^1 \\ \vdots & \vdots & \vdots & \vdots \\ p_1^{mn} & p_2^{mn} & \dots & p_N^{mn} \end{pmatrix}. \quad (1)$$

A linear mapping A from one video frame to an another in the sequence is assumed as these frames are correlated.

$$\begin{aligned} \mathbf{P} = [p_1, Ap_1, A^2p_1, A^3p_1, \cdots, A^{N-1}p_1]. \\ [p_2, p_3, \cdots, p_N] = A[p_1, p_2, \cdots, p_{N-1}]. \\ P_2 = AP_1. \end{aligned} \quad (2)$$

The system A is unknown and it captures the overall visual dynamics within the video. Computing eigen solution for A is computationally expensive since, A is large($mn \times mn$). For this reason, methods based on Krylov subspaces [39], [40], [41] could be efficiently used. The columns in the Krylov subspace are non-orthogonal but orthogonality could be achieved using Arnoldi iteration.

Equation 2 formulates the video frames in the span of krylov sequences. As a starting point of the Arnoldi algorithm, we have.

$$AP_1 \approx P_1 H. \quad (3)$$

Here, H a companion matrix is introduced that simply shifts from frames 1 through $N-1$ and approximates the last frame N by linearly combining the previous $N-1$ frames, i.e., $P_2 = c_0p_1 + ... + c_Np_{N-1} = \{p_1, p_2, p_3, \cdots, p_{N-1}\}c$.

$$H = \begin{pmatrix} 0 & 0 & \dots & 0 & -c_0 \\ 1 & 0 & \dots & 0 & -c_1 \\ & \ddots & \ddots & \vdots & \vdots \\ & & 1 & 0 & -c_{N-2} \\ & & & 1 & -c_{N-1} \end{pmatrix}. \quad (4)$$

In matrix form, we then have:

$$P_2 \approx P_1 H. \quad (5)$$

From Equations 3 and 5, we have

$$AP_1 \approx P_2 \approx P_1 H. \quad (6)$$

Hence this procedure will result in the low dimensional system matrix H. We solve the H matrix problem using eigenvalue analysis and obtain eigenvalues and vectors. It is known that eigenvalues of H approximate some of the eigenvalues of the full system A. The associated eigenvectors of H provide the coefficients for the linear combination that is necessary to express the dynamics within the video frame basis. H is calculated as follows:

$$H = U^{-1} L^{+} P_2. \tag{7}$$

Here, L^{+} is the pseudo inverse of L from the LU-decomposition of P_1 subspace (refer to the DMD processing steps in Algorithm 1).

Algorithm 1 Dynamic Mode Decomposition

```
**Input:** Sequence of images in a video $\mathbf{P} = [p_1, p_2, p_3, \cdots, p_N]$
**Output:** Dynamic mode DM corresponding to the phase angle = 0
$P_1 \leftarrow [p_1, p_2, p_3, \cdots, p_{N-1}]$
$P_2 \leftarrow [p_2, p_3, p_4, \cdots, p_N]$
$[L\ U] \leftarrow \mathbf{lu}(P_1)$
$H \leftarrow (U^{-1} L^{+} P_2)$
$[Evec\ Eval] \leftarrow \mathbf{eig}(H)$
$\lambda_j \leftarrow \mathbf{sort}(\mathbf{angle}(Eval_{jj}))$ # sort Evec correspondingly
$DM \leftarrow \mathbf{abs}(P_1 \times Evec)$ # project $P_1$ on the eigenvectors of $H$
$DMD\_mode\_select \leftarrow (DM(:,k))$ # $K$ is the index of the eigenvector,
    whose corresponding eigenvalue's phase angle is equals to or closest to 0
```

Dynamic modes not only contain the information about dynamic structures, but also about the temporal evolution of patterns within a video sequence. Since at no stage of the algorithm the system matrix A is needed, various extensions of the algorithm are possible. No specific spatial arrangement of the image data is assumed and there are no parameters to be tuned. The first dynamic mode captures the largest scale dynamics that is present in the sequence of frames [12], [13], [14], [15], [16], [17].

B. Principle Component Analysis (PCA) via snapshot approach

In order to compare DMD, this section explains how PCA can be calculated, and the output of which can be directly used to replace the DMD output. Therefore, we in here use a variant of PCA that is based on the 'snapshots' approach which is introduced by Sirovich [43]. Here each video frame at a particular interval in time is considered to be one snapshot. This method was introduced to reduce the computations of covariance matrices. Normally, computing PCA requires solving a $mn \times mn$ eigenvalue problem, where mn is the total number of pixels in a video frame. The snapshots method proposes that the covariance matrix can be approximated by a linear combination of the 'snapshots'. Instead of computing $mn \times mn$ covariance matrix, this method computes a covariance matrix of size $N \times N$. This method has become extremely popular in the field of CFD [44], [45], [46].

To implement the snapshots method, a sequence of images is converted to form a matrix as shown in Equation 1

The fluctuating pixel intensity matrix U is calculated by subtracting the mean matrix (average pixel intensities) $\hat{\mathbf{P}}$ from $\mathbf{P}$,

$$U = \mathbf{P} - \hat{\mathbf{P}}. \tag{8}$$

The covariance matrix is computed as

$$C = U^{T} U. \tag{9}$$

An eigenvalue problem for the covariance matrix is then solved

$$C A^{i} = \lambda^{i} A^{i}. \tag{10}$$

The eigenvectors are arranged according to the decreasing order of eigenvalues. This ordering is physically meaningful because, it reflects the fluctuating intensity energies in the PCA modes. Using the ordered eigenvectors the PCA modes are constructed.

$$\phi_i = \frac{\sum_{n=1}^{N} A_n^i U^n}{\| \sum_{n=1}^{N} A_n^i U^n \|}, \qquad i = 1, 2, \cdots, N. \tag{11}$$

In our experiments, we consider only the first PCA mode as this mode captures the greatest percentage of the intensity fluctuation from the video frames [16], [45], [47], [44].

C. Local Binary Patterns (LBP)

Local Binary Patterns (LBP), a powerful image representation, extracts texture information that is invariant to local gray-scale variations. During the LBP operation, every image pixel acts as a threshold to its neighbours to obtain a binary bit string which then forms a round number. In real applications, an image is usually divided into several small non-overlapping blocks of the same size to keep the spatial relation of objects. Conventionally, LBP is denoted as B-$\mathbf{LBP}^{u2}_{P,R}$, where P and R denote the neighbhourhood pixels and radius respectively and the superscript $u2$ denotes uniform LBP. B denotes the number of non-overlapping blocks. In this study we consider P = 8 for block-size = $\{1\times 1, 3\times 3, 5\times 5\}$ and radii = $\{1, 2, 3\}$. The optimal LBP settings are determined by cross validation.

D. SVM classifier - Histogram Intersection Kernel

Kernel based Support Vector Machine (SVM) is a widely used pattern classification method and is well known for its high classification accuracy. Kernel methods transform data from a low dimensional space to a high dimensional space using non-linear maps. By non-linearly mapping the data onto the high dimensional space, it is theoretically shown that any linearly non-separable data can become linearly separable [48], [49]. Since only the inner product between a pair of observations is required in the SVM formulation, the non-linear mapping does not need to be explicitly defined for an individual observation in the training set. Instead, the non-linear kernel function is also defined for a pair of observations. This maintains the efficiency of the SVM. Therefore, a kernel based SVM in general has a better performance over an original SVM for linearly non-separable classification tasks [50].

In this study we opted to use the Histogram Intersection Kernel [51], which is also known as the Min Kernel because it has been shown to be efficient and effective in image classification tasks [52]. The kernel is defined as follows:

$$K(x, y) = \sum_{i=1}^{n} min(x_i, y_i), \tag{12}$$

where n is the number of dimensions in the LBP histogram and x and y are the LBP histograms. Note that this kernel has no additional parameter to tune, which is convenient to use.

E. Performance and Threshold Criteria

Although the output of SVM can be used to make a hard decision, we shall use its soft output, which is defined as the distance of a test sample from the SVM decision hyperplane. For all experiments, we label samples derived from valid videos as positive; and those derived from attack videos to be negative.

Let $\mathcal{T}$ be the domain of SVM output. The decision is made by comparing the SVM output $t \in \mathcal{T}$ with a decision threshold, $\Delta \in \mathcal{T}$, as follows:

$$\text{decision}(t) = \begin{cases} \text{valid access} & \text{if } t > \Delta \\ \text{spoof/attack} & \text{otherwise,} \end{cases} \tag{13}$$

In order to calculate errors, we introduce match and non-match score sets, $\mathcal{Y}_1 \subset \mathcal{Y}$ and $\mathcal{Y}_0 \subset \mathcal{Y}$, respectively. This can result in two errors, namely, false rejection and false acceptance, the rates of which are calculated as follows:

$$\begin{aligned} \text{FRR}(\Delta) &\equiv P(t < \Delta|\omega_1) && (14) \\ &\approx \frac{|y|y \in \mathcal{Y}_1, y < \Delta|}{|\mathcal{Y}_1|} && (15) \end{aligned}$$

$$\begin{aligned} \text{FAR}(\Delta) &\equiv 1 - P(t < \Delta|\omega_0), && (16) \\ &\approx \frac{|y|y \in \mathcal{Y}_1, y < \Delta|}{|\mathcal{Y}_1|} && (17) \end{aligned}$$

respectively, where the conditioning variable ω_1 indicates that the comparison is due to a positive class (valid access) whereas ω_0 indicates that the comparison is due to a negative class (spoof/attack). The average of these both errors is defined by HTER.

$$\text{HTER}(\Delta) = \frac{1}{2}\left(\text{FAR}(\Delta) + \text{FRR}(\Delta)\right)$$

By taking the geometric average of FAR and FRR, HTER has the advantage that it is not affected by the overwhelmingly large sample size of one class versus another because both types of errors are weighted equally, thus coercing equal contribution of both errors, i.e., assuming equal prior probability for both classes. In our case, this is particularly desirable because the prior probability of an attack is difficult to estimate in practice. In the absence of any additional information, using equal prior probabilities is a reasonable option.

In order to report unbiased performance, the decision threshold that attains the optimal performance on a development set, i.e., Δ_* is then applied to the test set.

IV. Datasets

In this section we discuss the datasets used in our study. We have chosen to use datasets containing print attack and replay attack spoof samples. These datasets consist of short video recordings for both valid-access and attack attempts.

A. Print-attack Dataset

Print-attack Dataset [35] is the first publicly available dataset that provides a precise protocol with training, development and test sets. It contains a short video of valid access and spoof attacks to 50 different identities. The spoof attack that is emphasised in this dataset is print attack only, whereby an impostor presents a printed photograph of the targeted identity in order to falsify the access to a face biometric authentication system. This dataset, includes two different scenarios: (i) controlled background (the background is uniform) and (ii) adverse background (a non-uniform background). These scenarios provide a valid simulation of the attack environment.

B. Replay-attack Dataset

To extend the print-attack dataset, Chingovska *et al.* [19] introduced a replay-attack dataset with more sophisticated attacks. To validate the attacks, the author categorised the attacks in three different scenarios:

- *Print attack*: printed photograph (same protocols as proposed in print-attack).
- *Mobile attack*: the impostor presents photos and videos taken with a mobile phone using the mobile screen.
- *Tablet attack*: the impostor presents a high resolution digital photos and videos using an iPad screen.

The replay-attack dataset consists of 200 videos of valid access (with 375 frames each), and 1000 videos of attack attempts (with 240 frames each). The dataset is divided into three partitions, namely development, training and testing set. The development set is used for estimating the threshold value and training set is used for estimating any model parameters. Each of these sets are generated by a video gallery of 15 clients for development and training, and 20 clients for testing. This means that the training and testing sets are disjoint and completely independent of each other.

C. CASIA-FASD dataset

Additionally, to validate our proposed pipeline, we consider the CASIA-FASD spoofing dataset [20], which is more challenging than the first two data sets. For instance, it contains printed 2D attack with perforated eyes in order to simulate eye blinking, known as a ’cut-photo’ attack. This data set contains print attacks (or warped photo), cut-photo attacks and reply attacks captured in three settings, namely low, normal, and high resolutions. In total, the dataset consists of 600 videos including 150 valid-access and 450 attacks. The training and testing sets include 240 and 360 videos respectively.

V. Experiments & Results

In this section we discuss the experimental procedures as well as the results. The objectives of our experiments are as follows:

1) **Qualitative assessment of dynamic information captured by DMD.** From the outset, we conjecture that DMD can capture liveness cues such as eye blinks, lips movements and facial dynamics. Therefore, for valid videos, we wish to examine whether these features are

indeed picked up by DMD. In the case of photo attacks, DMD modes should, in theory, be characterised by a lack of motion, resulting in a somewhat noisy image. Finally, we expect that the facial texture properties of valid accesses and video attacks would differ due to the replayed videos being affected by the moiring or planar effects. Since DMD captures spatial-temporal features, it is reasonable to expect that the DMD modes would produce somewhat different textures.

2) **Optimisation of the DMD+LBP+SVM pipeline.** In our proposed DMD classification pipeline, there are two parameters that need to be pre-determined, namely, (1) the choice of DMD mode, and (2) parameters associated with the LBP parameters.

   Optimal DMD mode: In theory, for a video with N frames, we obtain $N-1$ DMD modes. Recall that each dynamic mode captures a "principal dynamics" axes of the video sequence. In order to select the most optimal mode, we choose the eigenvector whose corresponding eigenvalue has a phase angle that is equal to or closest to 0. Although it is possible to use all the dynamic modes, we limit our choice to a single mode because in our preliminary experiments, this configuration produced already very satisfactory classification performance.

   Optimal LBP parameters: For the LBP, we have to determine the optimal block size from the configurations $\{1\times1, 3\times3, 5\times5\}$ and the radius from $\{1, 2, 3\}$ pixels (in the DMD domain). The optimal block size and radius are determined experimentally.

   For the SVM, the histogram intersection kernel does not have any parameter to tune; and we shall use the standard SVM parameters without modification.

3) **Comparison of DMD with the state-of-the-art methods.** Once the optimal parameters have been determined on the development set, following the published Print Attack and Replay Attack protocols, the DMD-LBP-SVM pipeline is then applied to the evaluation (test) set. The interest in this experiment is to benchmark the generalisation performance of DMD against the state-of-the-art methods.

4) **Effect of window size on the system performance.** It is of interest to find out the impact of video window size on the generalisation performance. For this reason, we consider the window sets of top X×10-frames from the video sequence, where X ranges from $1, 2\ldots, 240$ (240 is the maximum number of available frames on the Print and Replay Attack databases). We conjecture that performance of DMD should improve with an increasing size of images available within a video.

5) **Additional research questions.** In the subsequent experiments, we shall also address three issues related to the peculiarities of the database and the DMD classification pipeline.

   a) **The impact of the face cropping on original images.** We note that in the replay-attack dataset, the faces in the attack videos are larger in comparison to the faces found in the valid videos. In order to investigate if the face size might influence the classification results, we consider only the face regions – as defined by the output of a face detector – with a background window of proportional size.

   b) **Pipeline with and without LBP features.** This experiment is designed to assess how much contribution is brought by LBP in the DMD+LBP+SVM pipeline. In theory, if LBP can successfully capture the video dynamics, one could simply do without DMD. In order to create this pipeline (without LBP), we directly use the histogram intensity of the chosen DMD mode – referred to as "DMD features" – and use the SVM with histogram intersection kernel as before. This results in a simplified pipeline labelled as DMD+SVM.

   c) **Comparison with PCA as a baseline method.** From the outset, supported by the literature in computation fluid dynamics, we claim that DMD gives better results than PCA for dynamic data. This experiment is, thus, designed to compare both methods. We note that this comparison is interesting in its own right because both methods are data-driven; and in addition, they share similar matrix decomposition algorithms, but differ only in the information being exploited. In particular, DMD is a motion-based method (thus capturing the dynamics in a video) whereas PCA is a nonmotion-based method. By comparing these two methods, we can gauge how much improvement is brought by the visual dynamics as captured by DMD; and as a result, how much degradation one might incur when this information is not exploited.

6) **Repeatability of experimental results on another data set.** As a final experiment, we shall also investigate the performance of the proposed pipeline on a different dataset i.e., the CASIA-FASD database. Not only is this database more challenging, it also contains another variant of the spoof attack produced by a 2D printed face photo with perforated eyes (cut photo), thus simulating eye-blinking. Since DMD is completely data-driven, we conjecture that DMD will capture the eye-blinking as well as the artefacts due to the printed face. This would present a challenging task for any facial liveness detection algorithm.

Finally, we conclude this experimental section by summarising the results of the overall experiments and by comparing with the existing state-of-the-art methods.

A. Potential of DMD

Figure 4 shows a valid video, a print attack video and a reply attack video, each recorded under either controlled or adverse conditions (in the top row). The middle and bottom rows show their corresponding first dynamic mode and PCA mode images. DMD captures the changes, fluctuations in intensities, and small variations obtained by suppressing the stable information present within the videos. Hence, in this case, DMD is able to suppress the background information

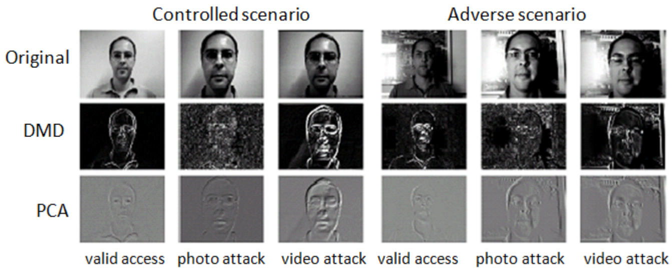

Fig. 4. Examples of the genuine and spoof attacks in controlled and adverse scenarios. The top row shows original images of valid access, photo and video attacks (left to right respectively). The middle row shows their corresponding first DMD mode and the bottom row shows their corresponding first PCA mode.

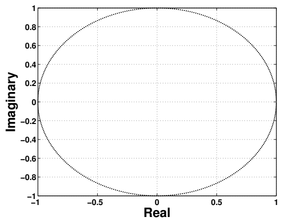

Fig. 5. Eigenvalue plot of DMD for one of the video clips in the replay-attack dataset. Each dot in the circle is a complex eigenvalue. If there are N images in the video, there are $N-1$ complex eigenvalues.

from the video sequence (which is stationary). DMD not only captured the movements in the facial expressions like eye blinks and lip movements but also the facial texture of the person inside the video frames. For the valid access, the texture of the person is weaker. On the other hand, we notice a stronger texture in the dynamic mode when the attack was carried out by a video.

It should be noted that the dynamic mode for the print attack displays a corrupted facial image. One reason is that a spoofing video with a printed photograph may have captured uniform illumination from the laptop screen with no dynamics inside the video. Another reason may be that there is no significant fluctuations within the intensities of this spoof video. As a result, the first DMD mode does not contain any distinctive movement.

In the case of controlled scenario the background was plain, with some shadings and shadows, as shown in the original videos. However, in the DMD mode, the background is completely suppressed. In addition, we observe that the dynamic mode for printed photograph attack contains a corrupted version of the individual's image with the background suppressed. These images show the strength of DMD as a preprocessing technique to distinguish between a valid and printed photograph attack.

B. Selection of DMD modes

For N number of frames, we obtain $N-1$ DMD modes. We calculate the phase angles based on the complex eigenvalues and select the eigenvector which has the eigenvalue phase angle = 0 (or the closest value = 0) and compute the dynamic mode.

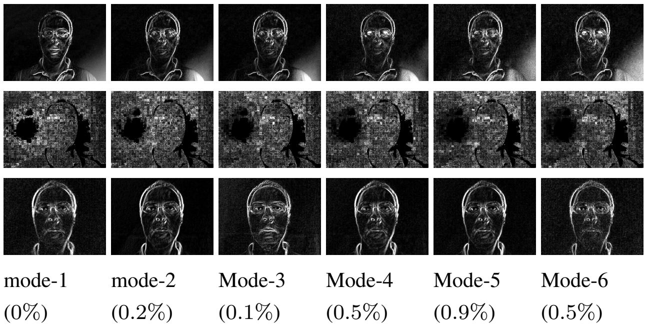
mode-1 (0%) mode-2 (0.2%) Mode-3 (0.1%) Mode-4 (0.5%) Mode-5 (0.9%) Mode-6 (0.5%)

Fig. 6. DMD modes from 1 to 6 (with HTER in % for test set on replay-attack dataset shown in brackets) showing the dynamics within the video frame sequence for valid access (top row), photo (middle row) and video attacks (bottom row).

Figure 5 shows the eigenvalues of upper Hessenberg matrix H, which represents the mapping between video frames. Unstable eigenvalues are located outside the circle [15] [14] and stable eigenvalues can be found on the circle [17] [16]. For the selection of modes, we calculate the phase angle for each of the eigenvalues, i.e, the eigenvalue at $(1,0)$ has phase angle = 0 and captures the overall dynamics from the video sequence. The phase angles above the axis $[(-1,0),(1,0)]$ correspond to positive phase and negative phase angles respectively. Each of these dynamic modes captures various dynamic information pertaining to the video sequence, from large scale to small scale, as shown in Figure 6.

Figure 6 shows the first 6 DMD modes of the three video samples of a valid video, photo attack and replay attack. These DMD modes are ordered in increasing phase angles with respect to the axis $[(-1,0),(1,0)]$. We observe that modes 4 and 6 of the valid video capture the complex dynamics of eye-blinking whereas modes 1 and 3 captures the lips movement of the video attack. For the print attack, DMD simply captures a non-uniform motion. The captured pattern does not have vitality signs such as eye-blinking, lips movement or head movement. The DMD modes for valid ones are observed with a different texture compared to video attacks. These properties may be sufficient to distinguish between a valid access and a spoof attack. We opted for dynamic mode-1 in our experiments as it gave us the best results by cross validation. This is supported by the literature [12], [13], [14], [15], [16], [17] which indicates that the first dynamic mode captures the largest dynamic scale or principal motion that is present in the sequence of images. In the next subsection, we assess the choice of LBP on the classification performance.

C. Impact of LBP parameters on the classification performance

1) On the replay-attack dataset: Recall that our implementation pipeline includes DMD, LBP and SVM (see Figure 3). In the training phase, the dynamic mode images corresponding to all the videos within the training set are given as an input to the LBP algorithm to obtain the LBP histogram features. The effectiveness of LBP depends on the different parameters for the block-size $B = \{1\times1, 3\times3, 5\times5\}$ and radius $R=\{1,2,3\}$. These parameters are optimised by minimising

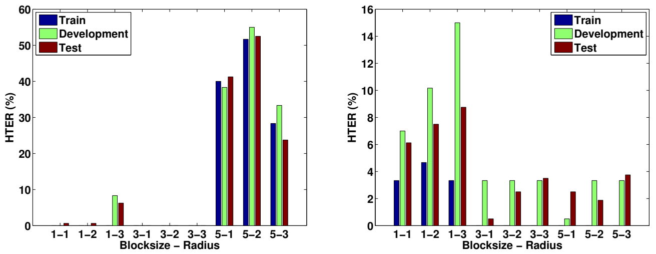

Fig. 7. Impact of block-size and radius of LBP on the HTER across the print-attack dataset (left) and replay-attack dataset (right).

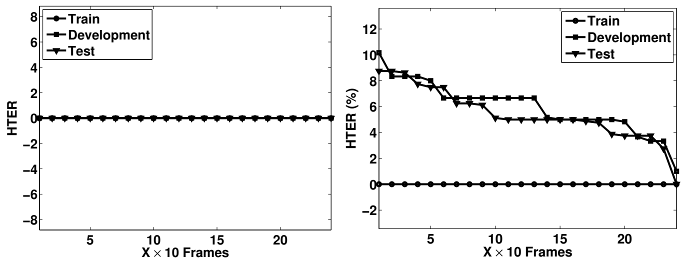

Fig. 8. HTER of the top X×10-frames evaluated on the development and test sets from print-attack dataset (left) and replay-attack dataset (right), here X ranges from $1, 2 \ldots, 240$.

HTER on the development set. We report our experiments on optimisation of these parameters.

On the training set, a HTER of 0% is attained across $B = \{1 \times 1, 3 \times 3, 5 \times 5\}$ and radius $R = \{1, 2, 3\}$ as shown in Figure 7 (right). For this reason, we opt to concatenate $3 - \mathbf{LBP}^{u2}_{8,1}$ and $5 - \mathbf{LBP}^{u2}_{8,1}$ features to form a larger feature vector. This hybrid feature vector can better capture the spatial arrangement of the texture properties within the video dynamics. The resultant LBP feature is of a dimensionality $1 - by - 2006$ (9×59 + 25×59; here 9 represents the 3×3 blocks, 25 represents the 5×5 blocks and 59 is the number of bins in the histogram). These concatenated LBP features for the train set are then fed to an SVM classifier with intersection kernel to train the classifier. We have not performed any parameter tuning on the SVM classifier but have selected 1 as the near boundary coefficient. We then proceed in a similar way to obtain the LBP features for videos in the development set and test set. The concatenated LBP features from the development set are fed into the trained SVM classifier to detect whether the input video is a spoof or not. At this stage, we use the class density distributions to obtain the threshold value = -0.1370. Later, this threshold value is used on the test set to discriminate between valid and spoofed videos.

2) On print-attack dataset: We proceed with a similar implementation process as in the case of the replay-attack in Section V-C1. Figure 7 (left), however, shows better results when compared to Figure 7 (right); in this, we observe an HTER of zero for $3 - \mathbf{LBP}^{u2}_{8,1}$, $3 - \mathbf{LBP}^{u2}_{8,2}$, $3 - \mathbf{LBP}^{u2}_{8,3}$ on the training, development and test sets respectively. Hence, it was sufficient to consider $3 - \mathbf{LBP}^{u2}_{8,1}$ for the print-attack dataset and no further concatenation of LBP codes was required.

Based on the development set, it is clear that $3 - \mathbf{LBP}^{u2}_{8,1}$ is a better choice for both print and replay-attack datasets. This turns out to be consistent with Chingovska *et al.*'s [19] work, although they worked in the spatial, static-image domain, whereas our method operates in the spatial, dynamic domain.

D. Performance at the video level

1) On replay-attack dataset: Using the threshold value = -0.1370 obtained on the development set, we compute the HTER score on the test set. The HTER scores of the propositions submitted to the 2^{nd} competition on "Counter Measures to 2D Face Spoofing Attacks" [54] competition were evaluated on the replay-attack dataset. The *CASIA* and *LNMIIT* teams achieved 0% HTER on both the development and the test set. Our method showed 0.5% HTER on development set and 0% on test set.

2) On print-attack dataset: The results for the print-attack at video level reveal a HTER values of 0%, 0% and 0% on training, development and test sets respectively. These results were also achieved by the *UOULU* and *IDIAP* teams in the 1^{st} competition on "Counter Measures to 2D Face Spoofing Attacks" [24] by using texture analysis methods.

E. Performance of window based approach

Since DMD needs to operate with a sequence of video frames, it is of interest to find out the minimum number of frames required. We conjecture that more frames should result in more stable performance; however, at the same time, we would like to know the minimum number of frames that is required to distinguish a valid access from attack videos. Consequently, we select the windows of the first $10, 20, \cdots 240$ frames from the training videos and evaluate it on development and test sets. We can consider at most 240 frames because all attack videos are truncated to 240 frames.

1) On replay-attack dataset: The first $10, 20, \cdots 240$ frames in the video sequence are considered and given as an input to the DMD algorithm. Therefore, for every training video in print and replay-attack datasets, we obtain 24 sets of windows, each of which differ by an increment of 10 frames. We then use the same LBP parameters as reported in Section V-C, to train each set using the SVM classifier and record the threshold value on the development set. The trained SVM is then applied to the test set to record the classification.

Figure 8 (right) shows HTER values for the first $10, 20, \cdots 240$ frames in the video sequence of the replay-attack respectively. We notice that the HTER on the train set was 0% in all the cases and the HTER on the development and the test set follows a decreasing trend as the number of frames in the train set are increased. Hence, this validates our conjecture that more frames result in a more stable performance. Although this is consistent with our expectation, the result also reveals that significantly more frames are required to detect replayed video attacks using tablets and mobile devices than the print attack (to be reported next)

The same experiment is repeated with the print attack and DMD attains 0% HTER, i.e., perfect classification. with just the first 10 frames across the training, development and tests as shown in Figure 8 (left). This result compares favourably with

TABLE III
RESULTS IN HTER $\%$ ON REPLAY AND PRINT-ATTACK DATASETS FOR DMD+SVM, DMD+LBP+SVM (FACE REGIONS AND ENTIRE VIDEO), PCA+SVM AND PCA+LBP+SVM (FACE REGIONS AND ENTIRE VIDEO) FEATURES

	Replay Attack		**Print attack**	
Pipelines	***Dev***	***Test***	***Dev***	***Test***
DMD+SVM (face region)	8.50	7.50	0.00	0.00
DMD+LBP+SVM (face region)	5.33	3.75	0.00	0.00
PCA+SVM (face region)	20.00	21.50	16.25	15.11
PCA+LBP (face region)	11.67	17.50	9.50	5.11
DMD+LBP+SVM (entire video)	0.50	0.00	0.00	0.00
PCA+LBP+SVM (entire video)	21.75	20.50	11.50	9.50

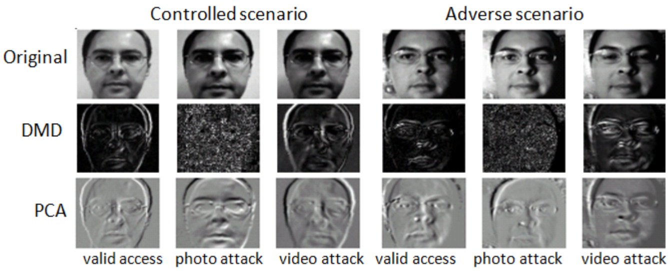

Fig. 9. Examples of the valid access and spoof attacks in controlled and adverse scenario for cropped face regions. The top row shows original images of valid access, photo and video attacks (left to right respectively). The middle row shows their corresponding first DMD mode and bottom row shows their corresponding first PCA mode.

Anjos *et al.*'s (2011) [18], in which the authors showed that 140 images are required to attain their best recognition performance. The method is based on optical flow. An enhanced version of their methodology in [35] requires more amount of images as an input for acheiving improved results. This suggests that DMD pipeline can capture dynamics information more efficiently than methods based on optical flow features.

F. Experiments on additional research questions

1) Performance on cropped frames: Due to the fact that spoof attacks in the replay-attack dataset have larger faces in comparison to the faces of the valid access, it is of interest to investigate if this might influence the classification results. Therefore we evaluate our experimental protocol only when the face regions are considered.

For this reason, we implement Viola-Jones face detection algorithm [55] to detect the face regions. All of the cropped images were of size 64×64. Examples of the valid access and spoof attacks in controlled and adverse scenarios for cropped face regions is shown in Figure 9. The top row shows original images of valid access, photo and video attacks (left to right respectively). The middle row shows their corresponding first DMD mode and bottom row shows their corresponding first PCA mode. Similar experimental protocol is followed and the results are shown in Table III. We see that an HTER of 5.3% was recorded on the development set of the replay-attack dataset and an HTER of 3.7% was recorded on the test set. The print attack recorded 0% on both the development and test sets.

2) DMD features vs *LBP features:* It is of interest to see whether LBP could bring any improvements to our proposed pipeline. Therefore, we investigate this setting by feeding DMD features directly to the SVM classifier. We do not investigate this scenario on the full images because the resultant image dimension will be prohibitively large (The full frame size is 240×320 and the resultant feature vector would be a size of $76800 \times V$, where V is the total number of videos). In the case of cropped images the feature vector is of size $4096 \times V$ since the frame size is 64×64. The results for DMD features recorded an HTER of 8.5% and 7.5% on development and test set respectively as shown in Table III. In the case of the print attack both development and test set have an HTER = 0. When compared to DMD+LBP features we see an HTER of 5.3% and 3.7% being recorded on the development and test sets of the replay-attack dataset. Therefore, we clearly see that the classification performance is improved using the LBP feature of the DMD+LBP+SVM pipeline.

3) DMD vs *PCA:* In order to further indicate the potoential significance of the DMD method we compare our results with the PCA algorithm as a baseline method. The results for PCA features give an HTER of 20% and 21.5% on the development and test sets of the replay-attack dataset respectively as shown in Table III. PCA+LBP features recorded an HTER of 11.6% and 17.5% on development and test set respectively. In the case of the print attack, both development and test sets have an HTER of 16.2% and 15.1% for PCA features while PCA+LBP recorded an HTER of 9.5% and 5.1% on development and test sets. Comparing to the results of DMD+SVM and DMD+LBP+SVM (entire video & face regions), PCA+SVM and PCA+LBP+SVM (entire video & face regions) show lower classification performance.

G. Proposed pipeline on CASIA-FASD Dataset

We have only considered the face regions for this dataset as the background of videos for different types of attacks are different, and this could possibly influence the classification performance. Examples of the valid and spoof attacks in CASIA-FASD (High Resolution HR) are shown in Figure 10. The top row shows original images of valid access, photo, cut photo and video attacks (left to right respectively). The middle row shows their corresponding first DMD mode and the bottom row shows their corresponding first PCA mode. The first DMD mode for the cut photo attack clearly shows high intensity values in the eye region as the eyes were continuously blinking for this type of attack.

Similar experimental protocols have been followed as above and the results are listed in Table IV. Once again, we observe that the DMD+LBP+SVM pipeline attains the best performance, although the error rates are considerably higher than those reported on the print and replay attack datasets.

H. Overall summary

In this study, we have demonstrated the potential of DMD as a preprocessing technique in highlighting the lip movements, facial textures and eye blinks across an entire video sequence. We have studied the effects of LBP parameters on the classification performance. Concatenation of the LBP codes $3-\mathbf{LBP}_{8,1}^{u2}$, $5-\mathbf{LBP}_{8,1}^{u2}$ to obtain an LBP code of

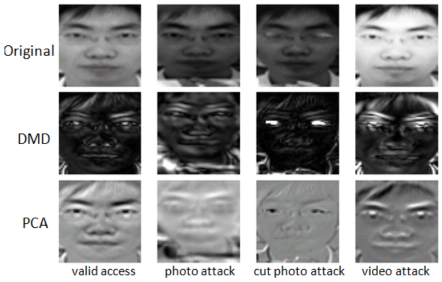

Fig. 10. Examples of the valid and spoof attacks in CASIA-FASD (High Resolution HR). The top row shows original images of valid access, photo, cut photo and video attacks (left to right respectively). The middle row shows their corresponding first DMD mode and bottom row shows their corresponding first PCA mode.

TABLE IV
CLASSIFICATION PERFORMANCE IN TERMS OF HTER (%) ON THE ON CASIA-FASD DATASET FOR DMD+SVM, DMD+LBP+SVM, PCA+SVM AND PCA+LBP+SVM

On face region	Test
DMD+SVM	29.50
DMD+LBP+SVM	21.75
PCA+SVM	33.50
PCA+LBP+SVM	24.50

dimensionality $1-by-2006$ performed well on the replay-attack dataset. For the print-attack dataset there is no need to concatenate LBP codes, since the codes corresponding to $3-\mathbf{LBP}_{8,1}^{u2}$, $3-\mathbf{LBP}_{8,2}^{u2}$ and $3-\mathbf{LBP}_{8,3}^{u2}$ demonstrated an HTER of 0% across training, development and test sets respectively. We have also conducted experiments at both the video level and frame level. At the video level, the proposed methodology achieves perfect recognition performance at HTER of 0% on the print attack data set; whereas for the replay attack, 0.5% and 0% is recorded on the development and test sets, respectively. These results show a good separability between the spoof and non-spoof samples for the considered datasets.

TABLE V
COMPARISON OF HTER (%) ON TEST SETS FOR THE PROPOSED PIPELINE WITH RESPECT TO THE CURRENT STATE OF THE ART. HERE, E: ON ENTIRE VIDEO AND F: ON FACE REGIONS.

	Algorithm	**Print attack**	**Replay attack**	**CASIA-FASD**
Anjos *et al.* (2013) [35]	Optical flow correlation	1.52	-	-
Schwartz *et al.* (2011) [31]	Partial least squares	-	-	-
Pereira *et al.* (2013) [56]	Motion correlation	-	11.79	30.33
	LBP	-	15.45	23.19
	LBP-TOP	-	8.51	23.75
		-		
Chingovska *et al.* (2012) [19]	LBP+LDA	-	13.87	
	LBP+SVM	-	18.17	
Proposed method	DMD+LBP+SVM^{E}	0.00	0.00	-
	DMD+LBP+SVM^{F}	0.00	3.75	21.75
	DMD+SVM^{F}	0.00	7.50	29.50
	PCA+SVM^{F}	15.11	21.50	33.50
	PCA+LBP+SVM^{F}	5.11	17.11	24.50
	PCA+LBP+SVM^{E}	9.50	20.50	-

Since DMD needs to operate with a sequence of frames in a video, it is desirable to find out the minimum number of frames required. For this purpose, we have repeated the experiments with progressively increasing frame sizes of $10, 20, \cdots 240$. Consistent with our conjecture, more frames result in more stable performance. For the print-attack dataset, our algorithm produces 0% HTER across all the datasets. This implies that the artefacts produced by the print attack videos do not contain any long-term dynamics; and that the attack can be detected by DMD effectively. In addition, we have addressed related research issues such as: i) evaluating the experimental protocol on face regions instead of evaluating on the entire video ii) comparing DMD with a baseline method i.e PCA, iii) addressing the importance of LBP in the methodological pipeline and finally evaluating the performance on a different dataset with diversified attacks. The experimental results show that: i) DMD+LBP+SVM pipeline on the face regions degrades slightly in performance when compared to the analysis on entire videos by HTER $3\% - 6\%$. ii) DMD based pipelines such as DMD+SVM and DMD+LBP+SVM work better than the PCA based pipelines, thus demonstrating the significance of motion based methods in distinguishing between valid accesses and spoof attacks compared to non-motion based methods such as PCA. Our experiments on CASIA-FASD dataset also show the strength of the proposed DMD+LBP+SVM pipeline. Our experimental results, following strictly the published experimental protocols, show that the proposed DMD pipeline achieved the best performance reported so far with an exception of komulainen *et al.*'s method [38]. The authors proposed to use context information such as facial, upper body and background information. In our case, we just used the facial features because our primary objective is to evaluate the effectiveness of DMD for distinguishing valid videos from the spoofed ones given only facial information. Komulainen *et al.*'s experiment informs us that using contextual information is extremely important. Indeed, when DMD is given the entire video frame instead of the face region only, the DMD+LBP+SVM pipeline also attains zero HTER (see Table V, proposed methods). This experimental setting is particularly interesting because the faces in the replayed videos have slightly larger size than those of the valid videos. This was designed to be so in order to avoid the frames of the replay device (tablet or mobile phone) from being captured and subsequently exploited by the algorithm designer. Nevertheless, DMD inadvertently captures the subtle cue in facial size difference, leading to perfect recognition on the test set. It is, therefore, reasonable to expect improved performance with DMD given more contextual information such as the upper body part as in komulainen *et al.*work. Our advantage, in our case is that the same DMD+LBP+SVM pipeline can be used straightforwardly without modification.

Finally, to conclude this summary section, we compare the experimental results obtained so far using variants of DMD with various methods reported in the literature. The results as indicated in Table V demonstate that DMD compares favourably, and even exceeds, current state-of-the-art methods.

VI. Conclusions & Future directions

This study shows the significance of the DMD method as a preprocessing technique when coupled with LBP and SVM to effectively detect spoof samples. We applied the DMD method to 1200 video clips of photo and video attacks on 50 clients, under different lighting conditions acquired from the replay-attack dataset; 400 video clips of photo attacks from the print-attack dataset; and 600 video clips from the CASIA-FASD dataset. The results are exceedingly promising in tackling the photo, cut photo and video attack challenges.

A. Conclusions

DMD can extract temporal dynamics efficiently in a data-driven manner and the resultant "texture dynamics" can be efficiently represented by LBP features with appropriate parametrisation. In the tested scenarios, $3 - \mathbf{LBP}^{u2}_{8,1}$ appears to be suitable across the varous print and replay attacks that we have tested on. This parameter is likely to be dependent on image resolution and camera quality. The SVM histogram intersection kernel works particularly well with the histogram features derived from LBP, with the additional convenience of no additional parameter tuning being required. The pipeline as a whole is thus efficient and easy to use, DMD and SVM in combination with the histogram intersection kernel hence requiring no tuning. Compared to other approaches reported to date, DMD appears to give superior performance. We attribute this to (1) the capacity of DMD to extract the dynamics of video sequence, as well as (2) the choice of pipeline within which DMD is deployed. Since DMD can capture video dynamics such as blinking eyes or mouth movements, or other non-obvious and subtle cues, this spares us from computing features that explicitly identify these cues. This is the main difference between our approach and cue-based methods. Although our work also makes use of LBP, it is used to represent the texture dynamics as opposed to static textures such as in the approach of Chingovska *et al.*'s [19]. We also note that a number of other methods have attempted to directly exploit dynamics such as those reported by Chakka *et al.* [24]. However, the performance reported in our work arises mainly from the capability of DMD to extract the dynamics of video automatically, and from the unique combination of DMD+LBP+SVM in a classification pipeline.

B. Future directions

Our locus of future research will aim to improve the DMD+LBP+SVM pipeline in the following ways:

- **Concatenation of all the dynamic modes.** As the DMD algorithm generates $N-1$ dynamic modes (where N is the number of frames in the video), it is of interest to concatenate some or all of these dynamic modes to represent a single image and produce corresponding LBP codes in order to further improve the classification performance, especially for very difficult video attacks.
- **Application of LBP on each dynamic mode**. LBP codes could be produced on each of the $N-1$ dynamic modes individually. It is possible to compute a mean LBP code from these modes in order to further improve the overall classifier performance.
- **Multiple kernel learning approach.** A linear combination of Mercer kernels is also a Mercer kernel. Consequently, a Histogram Intersection kernel could be computed for each of the dynamic modes individually and a linear combination of these kernels could then be considered for classification.

Acknowledgement

The funding for this work has been provided by Faculty of Engineering and Physical Sciences (FEPS), University of Surrey. Dr Poh also received funding from the BEAT project (grant no. 284989) to carry out this work.

References

[1] X. H. Li, Y. Q. Zhao, M. Liao, F. Y. Shih, and Y. Q. Shi, "Detection of tampered region for jpeg images by using mode-based first digit features," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–10, 2012.

[2] C. Roberts, "Biometric attack vectors and defences," *Computers & Security*, vol. 26, no. 1, pp. 14–25, 2007.

[3] N. K. Ratha, J. H. Connell, and R. M. Bolle, "An analysis of minutiae matching strength," in *Audio-and Video-Based Biometric Person Authentication*, pp. 223–228, Springer, 2001.

[4] K. Kollreider, H. Fronthaler, and J. Bigun, "Evaluating liveness by face images and the structure tensor," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, pp. 75–80, IEEE, 2005.

[5] A. da Silva Pinto, H. Pedrini, W. Schwartz, and A. Rocha, "Video-based face spoofing detection through visual rhythm analysis," pp. 221–228, IEEE, 2012.

[6] R. Lazarick, "Presentation attack detection," tech. rep., SC37, 2012.

[7] H. Meyer, "Six biometric devices point the finger at security," *Computers & Security*, vol. 17, no. 5, pp. 410–411, 1998.

[8] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial gummy fingers on fingerprint systems," in *Electronic Imaging 2002*, pp. 275–289, International Society for Optics and Photonics, 2002.

[9] L. Thalheim, J. Krissler, and P.-M. Ziegler, "Body check: biometric access protection devices and their programs put to the test," *ct*, vol. 11, p. 114ff, 2002.

[10] N. Duc and B. Minh, "Your face is not your password," in *Black Hat Conference*, vol. 1, 2009.

[11] A. Obied, "How to attack biometric systems in your spare time," *Internet: http://ahmed. obied. net/research/papers/biometric. pdf*, 2006.

[12] G. Tadmor, O. Lehmann, B. R. Noack, L. Cordier, J. Delville, J.-P. Bonnet, and M. Morzyński, "Reduced-order models for closed-loop wake control," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1940, pp. 1513–1524, 2011.

[13] M. Ghommem, M. Presho, V. M. Calo, and Y. Efendiev, "Mode decomposition methods for flows in high-contrast porous media. global–local approach," *Journal of Computational Physics*, vol. 253, pp. 226–238, 2013.

[14] P. J. Schmid, K. E. Meyer, and O. Pust, "Dynamic mode decomposition and proper orthogonal decomposition of flow in a lid-driven cylindrical cavity," in *8th International Symposium on Particle Image Velocimetry*, pp. 25–28, 2009.

[15] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, 2010.

[16] S. Tirunagari, V. Vuorinen, O. Kaario, and M. Larmi, "Analysis of proper orthogonal decomposition and dynamic mode decomposition on les of subsonic jets," *CSI Journal of Computing*, vol. 1, pp. 20–26, 2012.

[17] P. Schmid, L. Li, M. Juniper, and O. Pust, "Applications of the dynamic mode decomposition," *Theoretical and Computational Fluid Dynamics*, vol. 25, no. 1-4, pp. 249–259, 2011.

[18] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–7, IEEE, 2011.

[19] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, pp. 1–7, IEEE, 2012.

[20] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, pp. 26–31, IEEE, 2012.

[21] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, vol. 161. SIAM, 1974.

[22] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[23] R. H. Bartels and G. H. Golub, "The simplex method of linear programming using lu decomposition," *Communications of the ACM*, vol. 12, no. 5, pp. 266–268, 1969.

[24] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, *et al.*, "Competition on counter measures to 2-d facial spoofing attacks," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–6, IEEE, 2011.

[25] K. Kollreider, H. Fronthaler, and J. Bigun, "Verifying liveness by multiple experts in face biometrics," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pp. 1–6, Ieee, 2008.

[26] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," in *Defense and Security*, pp. 296–303, International Society for Optics and Photonics, 2004.

[27] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision–ECCV 2010*, pp. 504–517, Springer, 2010.

[28] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 3557–3560, IEEE, 2011.

[29] J. Maatta, A. Hadid, and M. Pietikainen, "Face spoofing detection from single images using micro-texture analysis," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–7, IEEE, 2011.

[30] R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, and F. Roli, "Fusion of multiple clues for photo-attack detection in face recognition systems," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–6, IEEE, 2011.

[31] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–8, IEEE, 2011.

[32] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*, pp. 233–236, IEEE, 2009.

[33] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[34] G. Pan, Z. Wu, and L. Sun, "Liveness detection for face recognition," *Recent advances in face recognition*, pp. 109–124, 2008.

[35] A. Anjos, M. M. Chakka, and S. Marcel, "Motion-based countermeasures to photo attacks in face recognition," *IET Biometrics*, vol. 3, no. 3, pp. 147–158, 2013.

[36] T. de Freitas Pereira, J. Komulainen, A. Anjos, J. M. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel, "Face liveness detection using dynamic texture," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 2, 2014.

[37] N. Poh, J. Kittler, and F. Alkoot, "A discriminative parametric approach to video-based score-level fusion for biometric authentication," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 2335–2338, IEEE, 2012.

[38] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pp. 1–8, IEEE, 2013.

[39] A. Krylov, "On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined," *Izvestija AN SSSR (News of Academy of Sciences of the USSR), Otdel. mat. i estest. nauk*, vol. 7, no. 4, pp. 491–539, 1931.

[40] Y. Saad, "Krylov subspace methods for solving large unsymmetric linear systems," *Mathematics of computation*, vol. 37, no. 155, pp. 105–126, 1981.

[41] A. Ruhe, "Rational krylov sequence methods for eigenvalue computation," *Linear Algebra and its Applications*, vol. 58, pp. 391–405, 1984.

[42] Y. Saad, *Overview of Krylov subspace methods with applications to control problems*. Research Institute for Advanced Computer Science, NASA Ames Research Center, 1989.

[43] L. Sirovich, "Turbulence and the dynamics of coherent structures. part i: Coherent structures," *Quarterly of applied mathematics*, vol. 45, no. 3, pp. 561–571, 1987.

[44] C. Duwig and P. Iudiciani, "Extended proper orthogonal decomposition for analysis of unsteady flames," *Flow, turbulence and combustion*, vol. 84, no. 1, pp. 25–47, 2010.

[45] V. Vuorinen, J. Yu, S. Tirunagari, O. Kaario, M. Larmi, C. Duwig, and B. Boersma, "Large-eddy simulation of highly underexpanded transient gas jets," *Physics of Fluids (1994-present)*, vol. 25, no. 1, p. 016101, 2013.

[46] K. Meyer, J. Pedersen, and O. Ozcan, "A turbulent jet in crossflow analysed with proper orthogonal decomposition," *Journal of Fluid Mechanics*, vol. 583, pp. 199–228, 2007.

[47] K. Meyer, D. Cavar, and J. Pedersen, "Pod as tool for comparison of piv and les data," in *7th International Symposium on Particle Image Velocimetry*, 2007.

[48] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[49] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[50] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[51] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. III–513, IEEE, 2003.

[52] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[53] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," tech. rep., DTIC Document, 1997.

[54] I. Chingovska, J. Yang, Z. Lei, D. Yi, S. Z. Li, O. Kahm, C. Glaser, N. Darner, A. Kuijper, A. Nouak, *et al.*, "The 2nd competition on counter measures to 2d face spoofing attacks.," in *ICB*, pp. 1–6, 2013.

[55] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[56] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?," in *Biometrics (ICB), 2013 International Conference on*, pp. 1–8, IEEE, 2013.

Santosh Tirunagari is a PhD student at University of Surrey, Department of Computing & CVSSP. He is responsible for developing algorithms for analysing sequential (longitudinal) data. More details about his research can be found at http://santosh-tirunagari.com/ and http://www.surrey.ac.uk/computing/research/msf/people/santosh_tirunagari/

Dr. Norman Poh joined the Department of Computing, University of Surrey as a Lecturer in Multimedia Security and Pattern Recognition in August 2012. He received the Ph.D. degree in computer science in 2006 from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. Prior to this appointment, he was a Research Fellow with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey. More details about his research can be found at http://www.ee.surrey.ac.uk/CVSSP.

Dr. David Windridge is Senior Lecturer in Computer Science at Middlesex University and heads the university's Data Science group. He is a visiting Professor at Trento University, Italy, and a visiting Senior Research Fellow at the University of Surrey. More details about his research can be found at http://personal.ee.surrey.ac.uk/Personal/D.Windridge/.

Aamo Iorliam is a PhD student at Department of Computing, University of Surrey. More details about his research can be found at http://www.surrey.ac.uk/computing/people/aamo_iorliam/.

Nik Suki is a PhD student at Department of Computing, University of Surrey. More details about her research can be found at http://www.surrey.ac.uk/computing/research/msf/people/nik_nurul_ain_binti_nik_suki/.

Professor Anthony T.S. Ho holds a Personal Chair in Multimedia Security and is currently Head of Department of Computing, University of Surrey since 2010. He also leads the Multimedia Security and Forensics research group in the Department. He is a Guest Professor of Wuhan University of Technology, China. More details about his research can be found at http://www.surrey.ac.uk/computing/people/anthony_ts_ho/.