A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

### Fake Video Detection with Talking Profile

By Wang Shu

Department of Computer Science School of Computing

National University of Singapore

AY 2018/2019

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

### Fake Video Detection with Talking Profile

By Wang Shu

Department of Computer Science School of Computing

National University of Singapore

AY 2018/2019

Advisor: Prof. Terence Sim Deliverables:

Report: 1 Volume

#### Abstract

With the rapid development of synthetic video generation and manipulation techniques, fake videos can now be easily produced even by amateurs. Fake videos could meddle with information reality in terrifying ways, and their appearances can be deceptive for the general public due to their semblance of realness. Politicians have become one of the most targeted group by fake video generators, as they frequently appear in public and their speeches are often very sensitive. Fake videos targeting politicians can become damaging weapons to erode trust in public officials and institutions, exacerbating social tensions and leading to genuine civic crisis. To defend people from internalizing fake videos, we need effective and robust authentication mechanisms. It is not sufficient, however, to only rely on the defects in faking techniques, which is bound to approach perfection over time. Instead, we describe here a biometric-based approach by establishing the talking profiles of the potential targets (e.g. politicians like Trump, Obama). Talking profile consists of multiple types of facial motions and can be viewed as the DNA of one individual’s talking behavior. Talking profiles can be extracted and established from the real videos of an individual, and then a classifier trained based on the talking profiles can be used to detect the unidentified fake videos. We also performed experiments to test the robustness of the talking profile against different video compression methods and shows that it is independent of video quality.

Subject Descriptors:

Computing methodologies - Biometrics Applied computing - Computer forensics

Computing methodologies - Machine learning approaches

Keywords:

Face Swap, Talking Profile, Machine Learning

Implementation Software and Hardware: Ubuntu 16.04, python 2.7.15

#### Acknowledgement

I would like to express my deepest appreciation to my advisor Prof. Terence Sim, my families, and friends. Without them, I would not have be able to complete this project.

# List of Figures

1.1 Face Swapping Illustration 3

* 1. Solution Pipeline 9
  2. Visualisation result of features extraction using OpenFace 10
  3. Eye landmarks 11
  4. Facial landmarks 11
  5. Illustration of Training/Testing Dataset Construction Process with *k* = 5 and

*m* = 2 13

**List of Tables**

**Table of Contents**

#### Title i

[Abstract ii](#_TOC_250017)

[Acknowledgement iii](#_TOC_250016)

[List of Figures iv](#_TOC_250015)

[List of Tables v](#_TOC_250014)

1. Introduction 1
   1. [Background 2](#_TOC_250013)
   2. [The Problem 3](#_TOC_250012)
   3. [Our Solution 4](#_TOC_250011)
   4. [Report Organization 5](#_TOC_250010)
2. Related Work 6
   1. [Facial Manipulation and Fake Video Generation 6](#_TOC_250009)
   2. [Fake Face Video Detection 7](#_TOC_250008)
   3. [Motion Based Face Recognition 8](#_TOC_250007)
3. Methodology 9
   1. [Talking Profile Construction 10](#_TOC_250006)
   2. [Dataset Generation 14](#_TOC_250005)
   3. [Classification Model 14](#_TOC_250004)
4. Evaluation 15
   1. [Experimental Setup 15](#_TOC_250003)
   2. [Results 15](#_TOC_250002)
5. Conclusion 16
   1. [Contributions 16](#_TOC_250001)
   2. [Future Work 16](#_TOC_250000)

**Chapter 1**

**Introduction**

The rapid development of synthetic video generation and manipulation techniques has enabled people to generate increasingly compelling fake videos at an incredibly low cost. Deepfake, for example, is one popular face-swap technique powered by deep learning. With few setup and training steps, President Obama can say whatever the creator wanted him to say in a photorealistic video.

Technology may be neutral, but the potential for malicious abuse is of grave concern, especially as the technology gets more refined. The fake videos generated using these techniques will undermine truth, confuse viewers, and sow discord at a much larger scale than we have already seen with text-based fake news.

To defend people from fake face videos generated by face-swap technologies, we need effective and robust authentication mechanisms. Instead of exploiting the defects of fake videos, here we describe a biometric-based approach by establishing talking profiles of the potential faking target (e.g. politicians like Trump, Obama). Talking profile consists of multiple types of facial motions and can be viewed as the ‘DNA’ of an individual’s talking behavior, and therefore independent of the visual quality of the fake videos.

### Background

In this section, we briefly discuss the history and background of the fake video problem. A detailed literature survey on some of the most important papers is presented in Chapter 2.

For many years, the creation of photorealistic synthetic videos is a laborious task, and requires strong expertise in computer graphics and animation. Therefore, synthetic video creation is mainly practiced by specialized filmmakers or visual effect producers. However, recent advances in fields of computer graphics( Zollh¨ofer *et al.*, 2018) and deep learning (Goodfellow *et al.*, 2014) have lowered the barrier of synthetic video implementation, and promised an unprecedented level of freedom in creative production such as synthetic performance (Chan *et al.*, 2018). On the other hand, these technologies also lead to controversial applications such as identity manipulation and falsification in images and videos. Due to the key role of faces in human identification and communication, many fake video generation techniques focus on facial manipulation. The landmarking, tracking and reconstruction of human faces have been well-examined in computer vision. In 2011, Dale *et al.* presented one of the first automatic face swap methods by constructing 3D model of source and target faces (Dale *et al.*, 2011). In an online article posted by Satya Mallick in 2016, he provided open source code to swap faces programmatically using OpenCV, with testifying illustrations of warping and

colour correcting Ted Cruzs’ face to fit that of Donald Trump (Mallick, 2016).

In late 2017, the generation of face-swap fake videos hit another major milestone when the Reddit user called ‘deepfakes’ published a series of face-swap fake videos in the forum. In 2018 Jan, a desktop software called FakeApp was developed and launched based on Googles TensorFlow framework as well as the original implementation from the Reddit user ‘deepfakes’. This App, understandably, further reduces the skill requisite of producing fake face videos, bringing deepfakes to the laymen and masses. On top of this, there are other implementations available, such as ‘deepfakes-faceswap’ on GitHub, but FakeApp remains the most accessible and popular due to its friendly interface.

In Figure 1.1, Cage’s face is overlaid onto the faces of multiple different actors. This example may seem at first entertaining and innocuous, but there are many malicious misuses of this technology, including the creation of fake news and fake celebrity pornographic videos. In 2019, some fake videos have gone viral and sparked off heated debate on the ethics of deepfakes and its AI



Figure 1.1: Face Swapping Illustration

backbone in general.

The most malignant applications of fake videos, however, would be those used to confound politics and international affairs. Face-swap techniques can be used to create precariously effective lies capable of inciting violence, discrediting leaders and institutions, or even tip the scale of elections. Several websites ban deepfakes, but the lack of reliable technology for detection makes accurately implementing this ban futile (AWS, 2018). Therefore, a precise, effective and robust algorithm to detect fake videos, especially those with face manipulation, is much in demand.

### The Problem

In this project, we targeted the fake videos generated with the most prevalent types of face manipulations techniques: face swapping. A character’s face in a video clip is swapped for another person’s face. Usually, the target under attack has his or her face placed on an impersonator or actor’s body. We aim to develop a biometric-based fake video detection algorithm, which could predict whether this face-swap has been performed for the concerned target in an unseen future video.

As mentioned earlier, there is a large incentive to target high profile politicians. We would as such focus on this character group. In this project, we require monocular videos that contain a person talking with clear facial

features.

### Our Solution

Existing detection algorithms mainly detect physical or statistical artefacts， relying on manually selected features or a deep neural network. (e.g. Li *et al.*, 2018). Instead of exploiting pixel-level characteristics of the video, here we describe a biometric-based approach by establishing talking profiles of the potential faking target. In the work by Zhang and Nejati *et al.* in 2013, they have shown experimentally that it is possible to distinguish similar-looking twins through acutely discerning their facial motion, which contains identity signatures (Zhang *et al.*, 2013). In face-swapped fake videos, the appearance of the face is changed to the target’s face, but facial and body motion remain unchanged, and thus do not belong to the target. Based on this observation, we propose our detection algorithm using talking profile which is composed of multiple types of facial motions to identify fake face videos. The pipeline of our method are as follows:

* + 1. Extract and establish the talking profile from videos. Talking profile contains multiple types of facial motions, such as head 3D rotation, eye-related movements and lip movements. The talking profile is represented as a sequence of motions for all frames of the video.
    2. An SVM classifier is trained for target character using the talking profiles extracted from videos of both the target and other persons. All videos used in this step are real and non-manipulated.
    3. For a new video of the target character, the classifier of that target character is used to predict the identity of the person in the video. If the identity is same as the target, the video is real; otherwise, the video is fake.

To carry out the experiments, we also generated face-swapped fake videos for Trump and Obama, by placing Trump/Obama’s face onto different characters.

### Report Organization

In chapter 2, a detailed literature review is presented. We survey on the existing facial manipulation techniques in fake video generation and detection methods. We will also review on works in the fields of motion-based face recognition. In chapter 3, we describe the detailed methodology of our algorithm. We present the experimental results and analysis in chapter 4. Chapter 5 is conclusion and discussion on future works.

**Chapter 2**

**Related Work**

### Facial Manipulation and Fake Video Generation

With the repertoire of fake video tactics that involve facial manipulation, there are two major categories, face swap and face re-enactment. As for the technical approach, both face swap and re-enactment can be realised via computer graphics based methods, such as subspace representation, geometry modelling (Lu *et al.*, 2016), or deep learning based synthesizing techniques.

#### Face Swap

For face swap, the face of person A in a video is replaced with another person B’s face, but the facial movement and expression of A will remain. This type of tactics has been used to insert celebrities’ faces into a variety of video clips in which they never appeared.

In the fields of computer vision and computer graphics, extensive research has been carried out in the areas of facial capture, tracking, animation and recognition, which lays foundation for various applications including face swapping. A comprehensive state-of-the-art report has been published by Zollh¨ofer *et al.* (2018).

Jones *et cal.* developed a real-time geometry capture approach to digital face replacement for a dynamic performance using high resolution performance geometry and textures captured by high speed cameras (2008). Dale *et al.* presented one of the first automatic face swap methods, which only requires single-camera video. They constructed a 3D multilinear model

to track the facial performance in both videos. Using the corresponding 3D geometry, they then warp the source to the target face and retime the source to match the target performance (2011). Garrido *et al.* presented a similar system that replaces the face of person A with the face of B while preserving the original expressions of person B (2014).

In recent years, a tremendous amount of work based on deep learning has been developed in face synthesis. The conception of Generative Adversarial Network (GAN) proposed by Good- fellow *et al.* in 2014 has since become a powerful framework for face generation. The research community’s effort on improving GANs for generative modelling has been prolific. Wasserstein GAN proposed by Arjovsky *et al.*(2017), for example, further improved the stability of training process. Liu and Tuzel proposed the coupled GAN (CoGAN) for learning a joint distribution of multi-domain images and applied on face images (2017). Later Liu, Breuel and Kautz proposed an unsupervised image-to-image translation framework based on CoGAN and achieved high quality image translation results on face image translation (2018).

#### Face Re-enactment

For face re-enactment, person A’s face will remain, but his or her facial movement and expression will be altered.

One of the first re-enactment methods has been proposed by Liu *et al.*. This approach transfers both expression and shading changes to a target neutral face of an actor (2001). Li *et al.* proposed a temporally more coherent image-based approach based on similarity metrics (2014). Face2Face proposed by Thies *et al.*, is an advanced real-time facial reenactment system, capable of altering facial movements in commodity video streams, e.g., videos from the internet. They combine 3D model reconstruction and image-based rendering techniques to generate their output (Thies *et al.*, 2016).

### Fake Face Video Detection

Comparing to the large amount of publications on GAN, there is a paucity of works on fake face video detection. Existing detection algorithms mainly focus on mesoscopic

properties of images and exploit physical or statistical image artefacts through manually selected features or a deep neural network.

Nguyen *et al.* applied capsule networks for replay attack, face-swap and face-reenactment detection(2018). These methods mostly focus on detection of fake face images; they consider videos as multiple independent frames and ignore the temporal information and correlation be- tween each frames. Guera and Delp, 2018, used a Recurrent Neural Networks (RNN) to capture the temporal inconsistencies between frames caused by face swapping. Matern *et al.* captured visual artefacts, such as missing shadow or reflection, missing face details, and inconsistency of eye colour to expose deepfakes and face manipulations (2019). Although these techniques detect a variety of fakes with relatively high accuracy, they may be affected by laundering counter- measures such as recompression, additive noise, which can easily destroy the measured artefact. Agarwal *et al.* proposed a technique that models facial movements typifying an individual’s speaking pattern. They used Pearson correlation to measure the linearity between the selected facial features in order to characterise an individual’s motion signature (2019).

### Motion Based Face Recognition

Psychological researches have shown that people rely on both facial appearance and motion for face recognition. Experiments (Hill *et al.*, 2001) have indicated that people use rigid head movement to identify faces.

Most of the motion based face recognition methods compute the displacement of key points on faces to represent facial motions. Tulyakov *et al.* computed the displacement of tracked points on faces and use them to identify different subjects (2007). Zhang *et al.* defined a talking profile which consists of six facial motions to distinguish twins, who have similar appearance but different talking pattern (2013).

## Chapter 3

**Methodology**

In this chapter we describe the construction of our detection algorithm.

At the training phase, we prepare two sets of videos, real videos of the target character *A* and *non-A* characters. From both sets of videos, we extract facial and head movements and establish the talking profiles, which are the positive/negative training set. We will then train a customised binary classier for character *A* with the training data. During testing phase, we prepare another two sets of videos, the real videos of the target character *A* and the fake videos of the target character *A*. We can then establish the positive/negative testing set with talking profiles. The binary classifier classifies a set of video frames to *A* (real) and *non-A*(fake). The pipeline of our approach is illustrated in Figure 3.1.

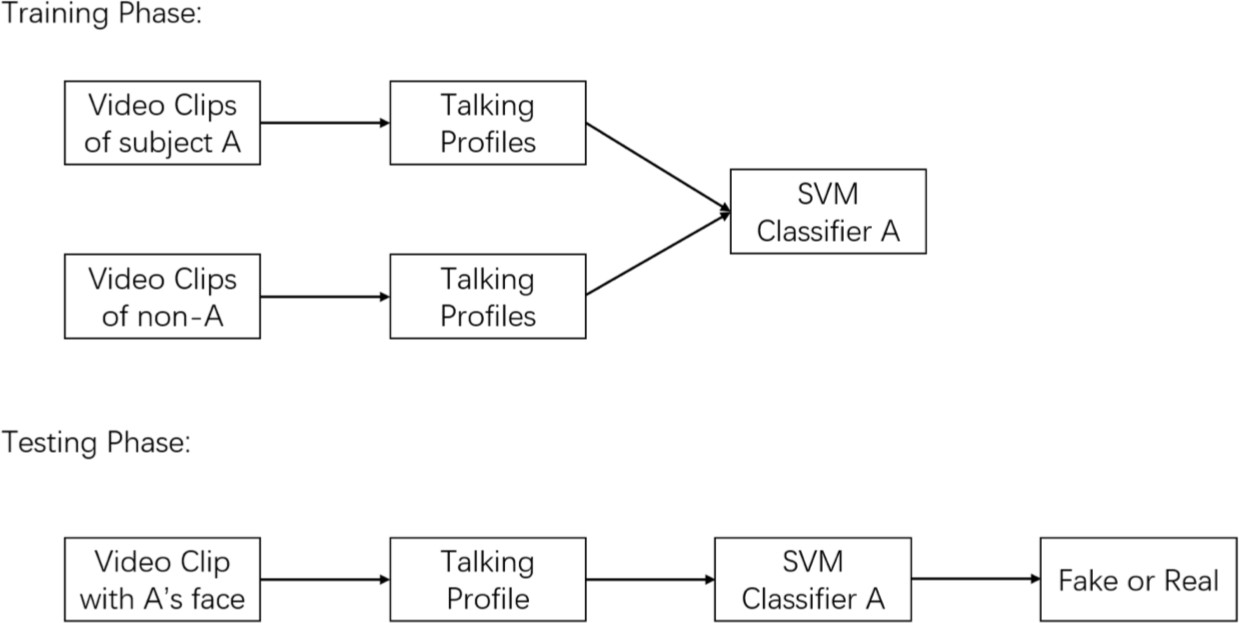


Figure 3.1: Solution Pipeline

### Talking Profile Construction

Talking profile of an individual is a continuous behaviour pattern, as such we need to first extract the facial motion states from each frame of the video, and then combine the static motion states of certain time interval (*k* frames) to build continuous features.

#### Facial Feature Extraction with OpenFace 2.0

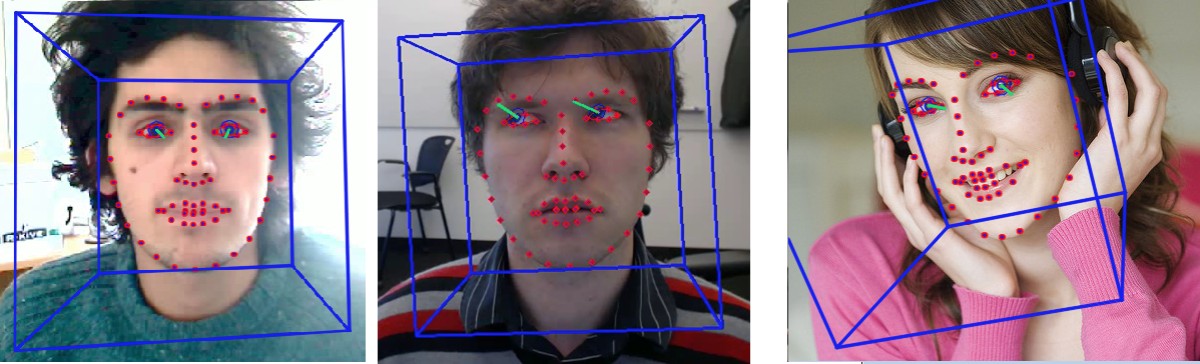
We use the open-source facial behaviour analysis library OpenFace 2.0: Facial Behavior Anal- ysis Toolkit (Baltruˇsaitis *et al.*, 2018) to extract facial and head movement from a video. OpenFace utilises constrained local neural fields(CLNF) (Baltruˇsaitis *et al.*, 2013) and con- volutional experts constrained local model for facial landmark detection and tracking (Zadeh *et al.*, 2017). Figure 3.2 illustrated the visualisation results of features extracted (Baltruˇsaitis *et al.*, 2018). For eye gaze tracking, it uses SynthesEyes dataset (Zhang *et al.*, 2015) to train the CLNF model and obtain eye gaze information (Baltruˇsaitis *et al.*, 2015). The facial action unit (FAU) occurrence detection and intensity estimation system is developed based on appear- ance (histograms of oriented gradients) and geometry features (shape parameters and landmark locations)(Baltruˇsaitis *et al.*, 2015). Figure 3.2 illustrated the visualisation results of features extracted, including facial landmarks, pose and gaze trackers (Baltruˇsaitis *et al.*, 2018).

Figure 3.2: Visualisation result of features extraction using OpenFace

For our model, we extracted and defined the following candidate facial features of each frame:

3-D Head location w.r.t Camera: *T˙* = (*tx, ty, tz*)

3-D head rotation: *R˙* = (*rx, ry, rz*)

Eye gaze direction: *tt˙* = (*lx, ly, lz, rx, ry, rz*) Normalised Eye Gap (blink indicator): *E˙* = ( *||p˙*11*−p˙*17*|| , ||p˙*39*−p˙*45*||* )

*lf ace lf ace*

Normalised Lip Gap/Width: *M˙*

= ( *||p˙*66*−p˙*62*|| , ||p˙*64*−p˙*60*||* )

*lf ace wf ace*

Both 3-D head location *T˙* and 3-D head rotation *R˙* are three-dimensional vectors, where positive

*tz* in *T˙* means the head is away from the camera and the head rotation *R˙* is in world coordinates

with camera being the origin. Eye gaze direction *tt˙*

have two components, which are left eye

gaze and right eye gaze, and both of them are three-dimensional vectors in world coordinates for respective eye. For eye gap *E˙*, we compute the distance between the upper and lower edge

points of eyes as illustrated in Figure 3.3 Eye landmarks.

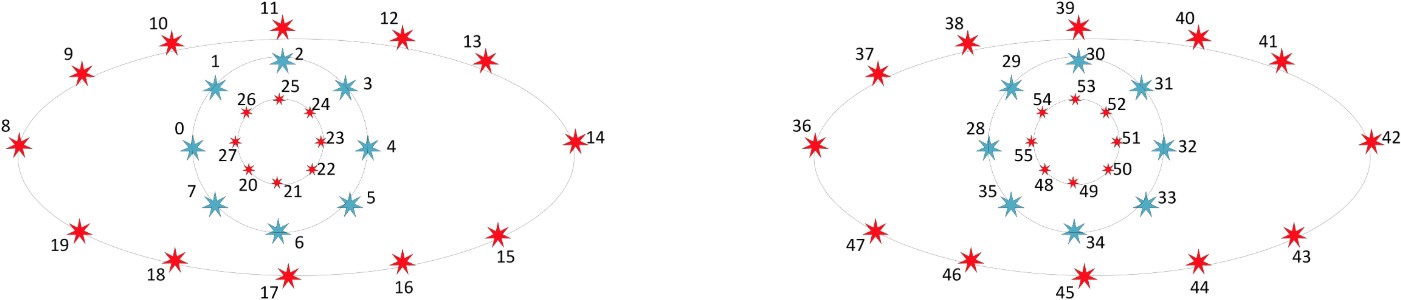


Figure 3.3: Eye landmarks

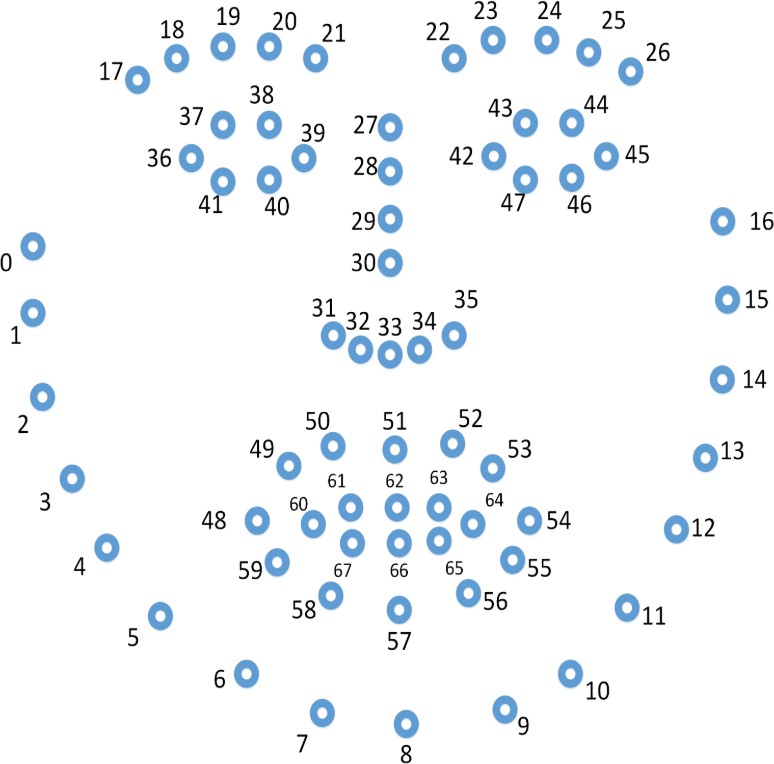


Figure 3.4: Facial landmarks

To measure mouth open-close magnitude *M˙* , we use the distance between the upper and

lower points of inner lip to represent the gap, and the distance between the left-most and right-most points of inner lip to represent the width. Figure 3.4 Facial landmarks illustrate

the landmark points extracted via OpenFace. For both *E˙* and *M˙* , we normalise the values by

dividing the length/width of the face, so that we could obtain the scale-invariant feature.

In addition, we will also include FAUs as candidate features. FAU encode the move- ments of facial muscles (Ekman, 1976). We constructed as vector with AU intensities *AU* = (*AUbrow, AUeye, AUnose, AUlips*), where:

*AUbrow* includes inner brow raiser (AU01), outer brow raiser (AU02),brow lower (AU04)

*AUeye* includes eye blink (AU45)

*AUnose* includes nose wrinkler (AU09)

*AUlip* includes upper lip raiser (AU10), lip corner puller (AU12), lip corner depressor (AU15), lip stretcher (AU20), lip tightener (AU23), lip part (AU25)

#### Talking Proftle

After extracting 6 types of facial features frame by frame, we can define the talking profile TP as a sequence of local motions between adjacent frames. Use *T Pi* start from *ith* frame. TP

starting at the 1*st* frame is constructed as the following:

 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *T˙*2 *− T˙*1 | *R˙*2 *− R˙*1 | *tt˙*2 *− tt˙*1 | *E˙*2 *− E˙*1 | *M˙* 2 *− M˙* 1 | *A˙U* 2 *− A˙U* 1 |
| *T˙*3 *− T˙*2 | *R˙*3 *− R˙*2 | *tt˙*3 *− tt˙*2 | *E˙*3 *− E˙*2 | *M˙* 3 *− M˙* 2 | *A˙U* 3 *− A˙U* 2 |
| : | : | : | : | : | : |

 

*T P*1 =

*T˙k − T˙k−*1

*R˙k − R˙k−*2

*tt˙k − tt˙k−*1

*E˙k − E˙k−*1

*M˙ k − M˙ k−*1

*A˙U k − A˙U k−*1

Where *T˙i, R˙i, tt˙i, E˙i, M˙ i, AUi* are 6 types of features extracted from frame *i* of the video. We consider *k* consecutive frames in the video each time, and we could vary the values of *k* to find optimal length. During implementation, we can concatenate each row of the matrix and obtain a flattened rray with length *lT P* = *lfeagure ×* (*k −* 1), where *lfeature* is the feature vector length of one frame.

#### Construct Training/Testing Dataset

*T P*1 consist of motion status from 1*st* frame to *kth* frame, and subsequently *T P*2 consist of motion status from 2*nd* frame to (*k* + 1)*th* frame, here *T P*1 and *T P*2 is overlapping by (*k −* 1) frames. In general *T Pi* and *T Pj* is overlapping by (*k −* (*j − i*)) frames. We may define the overlapping frames between two TPs to be *m*, then given the first TP is *T P*1, the next TP to be consider would be *T Pk*+1*−m*, starting from the (*k* + 1 *− m*)*th* frame and so on. We can then stack the stepped TPs to form our training/testing dataset.

For *K* total frames, with the consecutive frame length of each TP to be *k* and overlapping

frames between TP to be *m*, we can construct a training/testing samples of size *K−m−*1 and

*| ∫*

*k−m*

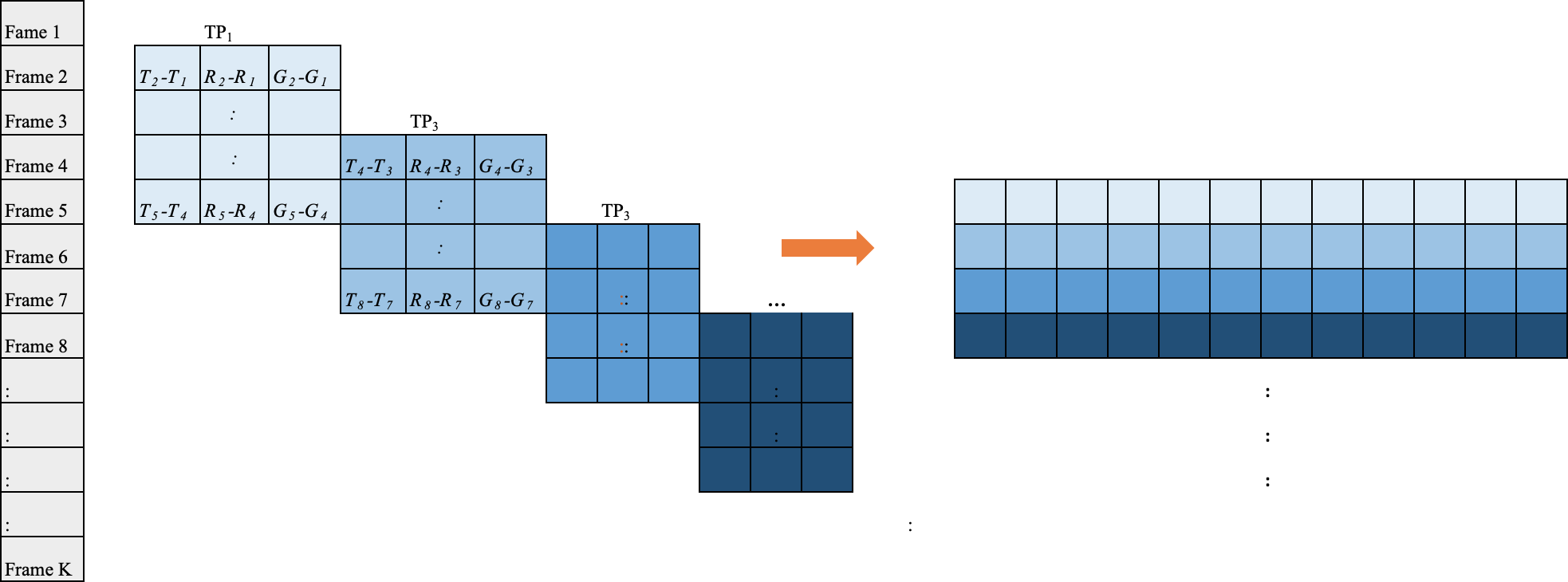
each sample of a vector of length *lT P* . Depending on whether the set of frames is from the real videos of the target character or other people, w can then label 1 or 0 for each row of the data frame. Figure 3.4 illustrate the construction process.

Figure 3.5: Illustration of Training/Testing Dataset Construction Process with *k* = 5 and *m* = 2

### Dataset Generation

#### FaceSwap

**Post-processing on Video Quality**

### Classification Model

**Training Testing**

**Chapter 4**

**Evaluation**

### Experimental Setup

### Results

**Chapter 5**

**Conclusion**

### Contributions

### Future Work