International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

# Anticipation of Forged Video Evidence using Machine Learning

**G. Abinaya[1], K. Sridevi Nattar[2], Dr. Rajini Girinath[3]**
[1, 2] UG Student, [3]Professor
Department of Computer Science and Engineering,
Sri Muthukumaran Institute of Technology, Chennai, India

## ABSTRACT

To detect audio manipulation in a pre recorded evidence videos by developing a synchronization verification algorithm to match the lip movements along with its audio pitch values. Audio video recognition has been considered as a key for speech recognition tasks when the audio is sullied, as well as visual recognition method used for speaker authentication in multispeaker scenarios. The primary aim of this paper is to point out the correspondence between the audio and video streams. Acquired audio feature sequences are processed with a Gaussian model. [1].This proposed method achieves parallel processing by effectively examining multiple videos at a time.In this paper, we train the machine by convolutional neural network (CNN) and deep neural network (DNN).CNN architecture maps both the modalities into a depiction space to evaluate the correspondence of audio –visual streams using the learned multimodal features. DNN is used as a discriminative model between the two modalities in order to concurrently distinguish between the correlated and uncorrelated components. The proposed architecture will deploy both spatial and temporal information jointly to effectively discover the correlation between temporal information for different modalities. We train a system by capturing the motion picture. This method achieves relative enhancement over 20% on the equal error rate and 7% on the average precision in comparison to the state of the art method.

***KEYWORDS:*** *Convolutional neural network, Audio-visual recognition, Deep neural network*

## I. INTRODUCTION

In this digital era, with highly sophisticated video editing tools, forgery videos can be submitted as evidence. Cyber forensics is a special field to manually verify the audio lip and speech synchronizations. Submitting a manipulated video is a serious offence, which may bias the result of judgment.The method of Audio Video Recognition systems is to influence the extracted information from one modality to surge the recognition ability of the other modality by complementing the missing data.Audio video speech recognition is a method that used image processing capabilities in lip reading to assist speech recognition systems in recognizing undeterministic phones or giving predominance among near probability decisions. The hardest part of an AVR algorithm is the feature selection for both the audio and visual modalities which in turn directly creates impact on the AVR task. The modules in the lip reading and speech recognition work individually and their results are merged at the stage of feature fusion.

In the area of machine learning, deep learning approaches have recently fascinated increasing attention because deep neural networks can effectively pull out robust latent features that enable various recognition algorithms to reveal revolutionary generalization capabilities under diverse application circumstances. According to the speech modality, speech recognition systems uses Hidden Markov Models (HMMs) to extract all the temporal information of speech and Gaussian mixture Models (GMMs) to separate between different HMMs states for acoustic input representation. A neural training algorithm is used to automatic the human intelligence

of detecting the non-synchronization between the video lip and voice. Deep learning has recently been engaged as a mechanism for unsupervised speech feature extraction. Deep learning is also used for extracting feature of unlabeled facial images. A queue based parallel synchronization verification architecture will be developed to make a log of mismatch time in video along with its video name or id.

Video features like shape, motion and audio features of amplitude and pitch frequency will be compared with the neural trained database to predict the authentication of the input video. The main aims to check if a stream of audio matches with a lip motion within the specific amount of time.

## II. METHODOLOGY

For starting this work, we primarily captured a video using normal mobile phone camera. The video is extracted in the format of MP4 and MOV with the audio format in MP3 and MPEG. The extracted video is then divided into number of frames. The changes in the motion of the object is detected, so that when the object in the video moves their motion can be traced. In this method, we take only short interval of video clip (0.3-0.5s). The ultimate aim of this paper is to find the correlation between audio and video. Sometimes voice we hear may differ from the visual we experience. Likewise, evidences submitted may be manipulated by any other audio clips and may collapse the judgment. Still in cyber forensics, no automation were invented to undergo testing of video evidence. Manual judgment may decrease the accuracy of video which in turn increases the crime. Hence to check the evidences submitted by a delinquent we need to deploy a method which checks automatically about the truthfulness of the evidence.

All platforms (OS) will support AVI, including uncompressed, indexed, greyscale, and Motion JPEG-encoded video (.avi) Motion JPEG 2000 (.mj2). All original versions of windows will support MPEG-1 (.mpg) Windows Media® Video (.wmv, .asf, .asx). The platform that is emerged after windows 7 will support MPEG-4, including H.264 encoded video (.mp4, .m4v) Apple QuickTime Movie (.mov).

Here we have invented a method to check the video evidence automatically that is being submitted. We have incorporated this method with machine learning in order to train our system. The submitted video undergoes several stages of inspection in small interval of time. The various stages includes face recognition, face detection, lip detection, capturing motion of video and at last the correspondence between audio and visual streams.

### LIP DETECTION:

### RGB APPROACH:

In RGB space,skin and lips have differnet components. The red is dominant for both, the green is more than the blue in the skin colour, and skin appears more yellow lips. Gomez et al., propose a method for lip detection in which the image is transformed by a linear combination of red, green and blue components of the RGB colour space. Then they apply a high pass filter to highlight the details of the lip in the transformed image, after which both of the generated images are converted to obtain a binary image. The largest area in the binary image is recognized as the lips area.

### HSV APPROACH:

The RGB colour system is most commonly used in practice, while the HSV is closer to how humans describe and interpret colour. Hue(H) represents the dominant colour as perceived by someone, apparently, hue is a key feature for detecting lips using HSV colour system, and since the hue value of the lip pixels is smaller than that of the face pixels. Coianix et al., use the filtering hue method to detect the lips in coloured images. [2] Researchers normally detect the face then use it to detect the lips, while Jun and Hua, proposed a reverse technique, by which they detect the face using the lips. Their method determines skin area based on a threshold value in an HSV colour model, and then it searches the rough lip area based on a threshold value in an HSV colour model, and then it searches the rough lip area based on the skin chrominance adaptively. After that it employs an RGB filter to find the lip, depending on the geometrical relation between face and lip; the face is finally located.

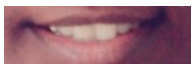

Fig 1: lip detection

Lip detection may vary from one video to another video. It is completely depends upon the angle that the video is captured. If the video is in different angle

than the angle which is being trained, machine itself will add or minus 90° from the video.

**POWER SPECTRAL DENSITY:**

It displays the strength of the deviations(energy) as a function of frequency. Power Spectral density(PSD)[3] shows at which frequency variations are strong and at which frequency variations are weak. Unit of PSD is energy per frequency. One an obtain energy within a specific frequency range by integrating PSD within that frequency range. PSD is computed directly by a method called FFT or computing auto correlation function and then transforming it. PSE is the spectral characteristics of signals characterized as random processes. The autocorrelation function of a random signal is the appropriate statistical average that will use for the characterizing random signals in the time domain, and the power density spectral is the Fourier transform of the autocorrelation function, provides the transformation from the time domain to the frequency domain.

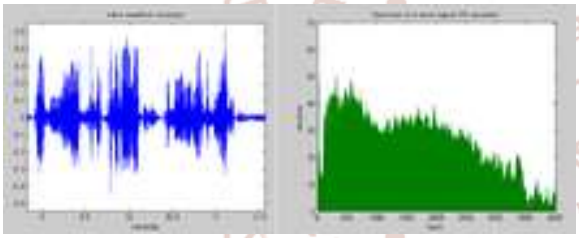

FIG 2 : voice waveform(left)
Power spectrum(right)

A signal is decomposed into a number of spectrum of frquency over a continuous range. PSD is ditribution of power into frequency components composing that signal. There are two methods of estimating PSD. They are parametric and non-parametric method. Parametric method finds the parameter for a mathematical model describing a signal,system or process.

Parametric methods are constructed on parametric models, such as AR models, moving average (MA) models, and autoregressive-moving average (ARMA) models. Parametric methods is otherwise termed as model-based methods. To estimate the PSD of a time series with parametric methods, first obtain the model parameters of the time series Parametric methods on the other hand can provide high resolution in addition to being computationally efficient. The most common parametric tactic is to arise the spectrum from the parameters of an autoregressive model of the signal [4]

Non parametric method is held out by estimation. Estimation is approximating output of a system after its degraded by noise or distortion. So for a random signal only an estimation can be obtained. In non parametric power spectrum estimation there are five methods. Periodogram method, Modified periodogram method, Barelett's method, Welch's method, Blackman-Tuckey method.

**WELCH'S METHOD:**

Welch's method is an improved method obtained by changing Bartlett method in two aspects. First, the data segments in the Welch method [5] are allowed to overlap. Second, each data segment is windowed prior to computing the periodogram. Blackman tuckey method can be used when the data length is short, but when the data length increases Welch method gives excellent results. The length of the applied window controls the trade-off between bias and variance of the resulting power spectral density (PSD).

The goal of spectral analysis is to decompose the data into a sum of weighted sinusoids. This decomposition allows one to assess the frequency content of the phenomenon under study. The phenomenon under study may be concentrated in some narrow frequency band. On the other hand, it might be spread across a broad range of frequencies. Spectral analysis is divided into two major areas. One can compute a Fourier transform or a power spectral density (PSD). When the data contains no random effects or noise, it is called deterministic. In this case, one computes a Fourier transform. One computes a PSD when random effects obscure the desired underlying phenomenon. To see the desired phenomenon, some sort of averaging or smoothing is employed.

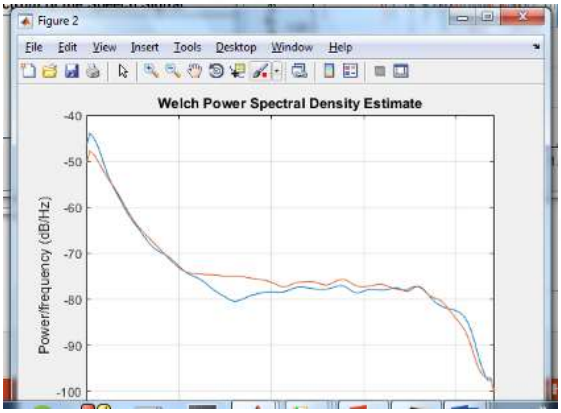


FIG 3:Welch PSD representation

## III. FEATURES OF AUDIO CLASSIFICATION

Five audio classes such as popular music, classical music, speech, noise and crowd noise are differentiated by calculating the ability of four derived audio features. [6] The feature sets include low-level signal properties, Mel-frequency spectral coefficients, and two new sets based on perceptual models of hearing. The temporal behaviour of the features is analysed and parameterized and these parameters are included as additional features. Using a standard Gaussian framework for classification, results show that the temporal behaviour of features is important for automatic audio classification. If the Audio classification is constructed on features from auditory perception than on standard features, then the result would be well. Hence in this paper we classify audio by auditory perception rather than standard audio sets. The audio is extracted and then graph is plotted to make 0's and 1's.

## IV. Machine learning techniques in Image proessing

Machine Learning is an interdisciplinary field involving programs that improve by experience. It is good for pattern recognition, object extraction and colour classification etc. problems in image processing problem domain. Machine learning constructs computer programs that develop solutions and improve with experience. It solves problems which cannot be solved by enumerative methods or calculus-based techniques. Machine learning is a part of Artificial Intelligence. Machines are trained with data sets. Machine learning objectives is to develop the computer algorithms which can acquire experience from example inputs and make data-driven estimates on unknown test data. Such algorithm can be divided into two types. They are supervised and unsupervised learning. They are complementary to each other. Meanwhile supervised learning forces labels of inputs which are significant to human, it is easy to apply this kind of learning algorithms to pattern classification and data regression problems.

## APPLICATION OF MACHINE LEARNING:

**Virtual Personal Assistants such as** Siri, Alexa, Google. **Predictions while Commuting includes** Traffic Predictions and Online Transportation Networks, **Videos Surveillance, Social Media Services such as sorting newsfeed according to user. Email Spam and Malware Filtering, Online Customer Support, Search Engine Result Refining, Product Recommendations, Online Fraud Detection.**

## V. DATA REPRESENTATION

The proposed system uses two non-identical ConvNets which uses a pair of speech and video streams.

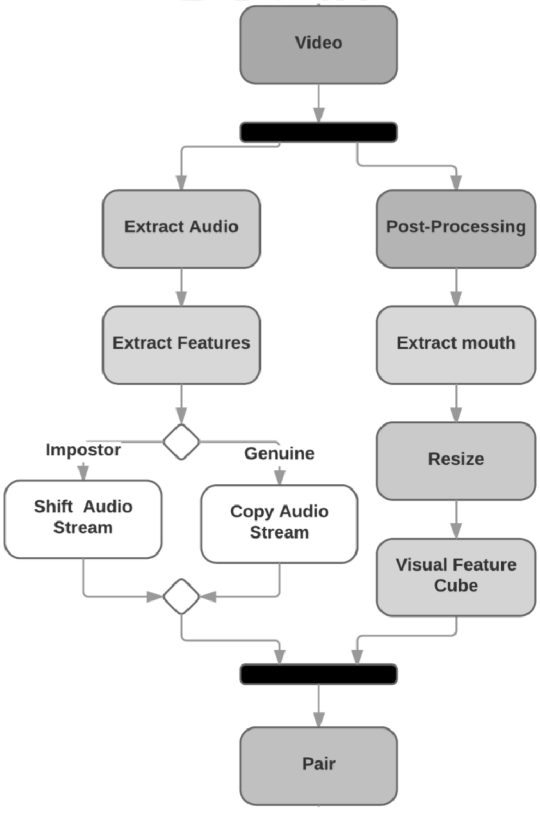


Fig4: Flow of video

The input is a couple of features that characterize lip movement and speech features extracted from 0.5 second of a video clip. The mission is to govern, if a stream of audio links with a lip motion clips within the desired stream period. The effort of this mission is the short time break of the video clip (0.3-0.5 second)

considered to assess the method. This setting is close to real-world situations because, in some biometrics or forensics applications, only a short amount of captured video or audio might be available to distinguish between different modalities. Temporal video and audio features must correspond over the time interval they cover.

A. SPEECH

Locality plays the main role in CNNs. The convolution operation is applied to specific local regions in an image. Since the input speech feature maps are treated must be locally correlated in the sense of time and frequency on both axes respectively.

B. VIDEO

This video clip uses frame rate of 30 f/s. Subsequently, 0.3 second visual streams constructs 9 frames. The input size of the visual stream of the network is of size 9 × 60 × 100, where 9 is the number of frames that represent the temporal information. An example of mouth area representation is provided in Fig5.

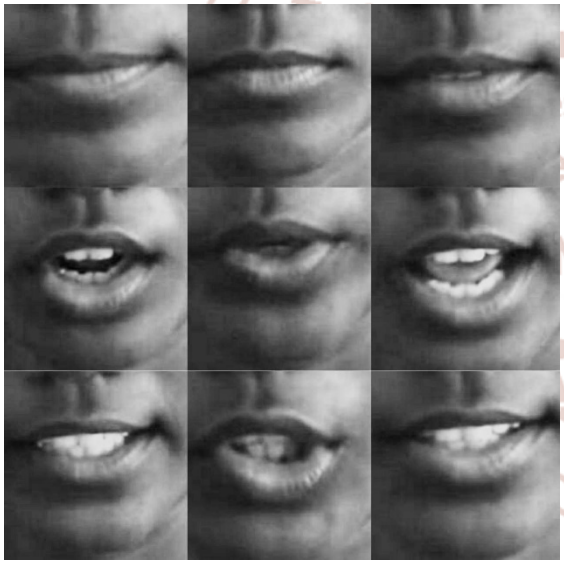

FIG 5 : The sequence of mouth areas in a 0.3-second video stream.

A relatively small mouth crop area is deliberately chosen due to practical concern because, in real-world circumstances, it is not very likely to have access to high resolution images. Moreover, unlike the usual experimental setups for CNNs, we did not restrict our experiments to input images with uniformly square aspect ratios.

## VI. CONCLUSION

The audio and video correspondence are tested with the plotted graph which is obtained by extracting the audio and video features. If the audio and video streams or patterns matches with each other, then the audio and video is synced properly. If the pattern does not match with each other then there arises a mis-sync in audio and video file. The unmatched interval of video file is retrieved and it is stored in the database to make juries to take decisions on the submitted video. This paper is also used to calculate the lag between the video motion and audio sequence. It can handle large datasets of file without time complexities. Log report of audio mis synchronization in particular time period will be automatically generated after analysing results. Hence this paper can be used to find the lip synchronization more accurately in order to avoid crimes that arises by submitting manipulated video.

## VII. REFERENCES


1) 3D Convolutional Neural Networks for cross audio-visual matching recognition, Amirsina Torfi, Syed Mehdi Iranmanesh, Nasser nasrabadi (Fellow IEEE) and Jeremy.
2) Color based lip localization method, Ahmad B.A Hassanat,Sabah Jassim,Applied computing,The University of Buckingam,UK
3) Analysis of Power Spectrum Estimation Using Welch Method for Various Window Techniques, Pranay Kumar Rahi , Rajesh Mehra , National Institute of Technical Teachers' Training &Research.
4) Emmanual C. Ifeachor, Barrie W. Jervis,"Digital Signal Processing A practical Approach, Person Education, Second Edition, pp. 864-868, 2005.)
5) PSD Computations using Welch's method,Otis M.Solomon, Jr,Division 2722,Test data analysis
6) Features of audio classification, Jeroen Breebaart and Martin McKinney, Philips Research Laboratories, Prof. Holstlaan
7) 3084–3097, Nov. 2015.