International Research Journal of Natural and Applied Sciences
ISSN: (2349-4077)
Impact Factor 5.46 Volume 4, Issue 1, January 2017
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

Comprehensive feature extraction for the recognition of visual speech and speakers

Mistry Kaushal K
Lecturer, Computer Engg. Department, B & B Institute of Technology, VVnagar

Authors Address:
Mistry Kaushal K
Lecturer,
Computer Engineering Department,
B & B Institute of Technology (BBIT), VVnagar, Gujarat

ABSTRACT

Visual speech recognition (VSR), commonly referred to as lip reading, along with visual speaker recognition (VSRec), which utilizes visual cues for identifying speakers, are rapidly advancing domains. These technologies find applications in areas such as assistive devices for individuals with hearing impairments and sophisticated security systems. The effectiveness of these systems is significantly influenced by the quality of the visual features extracted. This paper provides an extensive review of various feature extraction methods employed in VSR and VSRec, classifying them into three main categories: appearance-based, geometric-based, and deep learning-based techniques. The advantages and limitations of each category are examined, as well as their appropriateness for particular applications. Additionally, this paper investigates current research directions in feature fusion and temporal modeling aimed at improving the robustness and precision of visual speech and speaker recognition systems.

1. INTRODUCTION

Humans inherently depend on visual signals to enhance auditory information during the process of speech perception, particularly in environments with significant background noise. This natural capability serves as the basis for Visual Speech Recognition (VSR), which seeks to transcribe spoken language solely from visual data. In a similar vein, subtle movements and expressions of the face can disclose distinctive traits of an individual, facilitating Visual Speech Recognition (VSRec).

Both VSR and VSRec encounter difficulties in accurately capturing pertinent visual data. Elements such as variations in lighting, changes in head orientation, differences among speakers, and the intrinsic ambiguity of visemes (the visual equivalents of phonemes) considerably affect their performance. Consequently, effective feature extraction is essential for addressing these challenges and developing robust and dependable systems. This paper intends to deliver a thorough overview of the various feature extraction techniques utilized in the domains of VSR and VSRec

2. Traditional Feature Extraction Techniques:

Traditional feature extraction methods can be generally divided into two main categories: appearance-based and geometric-based approaches.

2.1 Appearance-Based Features:

Appearance-based features make use of pixel intensities or derived representations from the mouth region. These features are relatively straightforward to implement and are capable of capturing comprehensive visual information.

Pixel Intensity Based Features: The raw pixel intensities within a designated region of interest (ROI) surrounding the mouth are utilized as features. Although this method is simple, it is vulnerable to noise and changes in lighting conditions. Common strategies to enhance robustness include normalization, resizing, and the application of dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA).

Discrete Cosine Transform (DCT): The DCT converts the spatial characteristics of the mouth region into the frequency domain. Typically, the lower frequency components, which represent the overall shape and structure, are preserved as features. This method provides a degree of resilience against high-frequency noise and variations in lighting.

Gabor Filters: Gabor filters serve the purpose of extracting features across various orientations and scales, effectively capturing texture details and edge information within the mouth area. They are particularly adept at detecting the dynamic alterations in the lips and adjacent regions.

Local Binary Patterns (LBP): LBP offers a robust and efficient method for characterizing the local texture of an image by assessing the intensity of each pixel in relation to its neighboring pixels. This technique is invariant to monotonic changes in grayscale, thereby providing a degree of resilience against variations in illumination.

2.2 Geometric-Based Features:

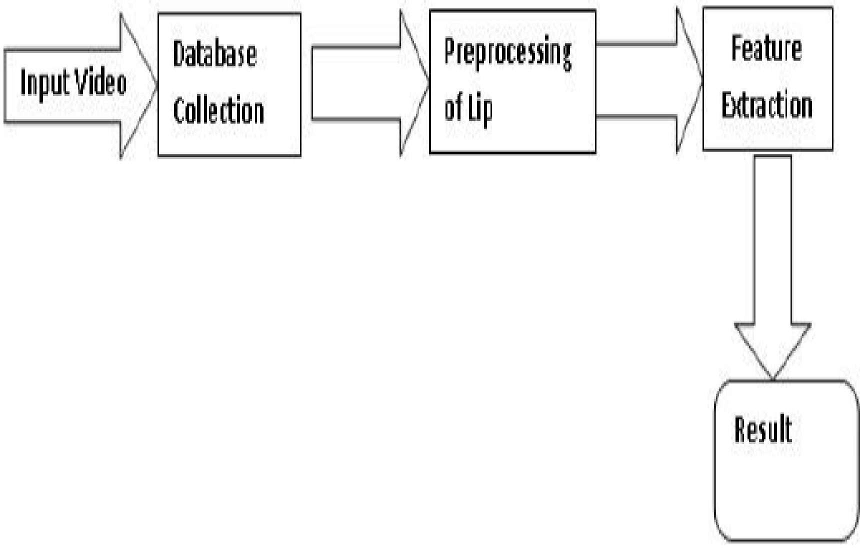
Geometric-based features focus on extracting distinctive points and contours from the mouth region, thereby representing its shape and movement. These features tend to be more resilient to changes in illumination compared to those based on appearance.

Active Appearance Models (AAMs): AAMs are statistical frameworks that integrate shape and texture information to depict the face or mouth region. They are developed using a comprehensive dataset of labeled images and can be employed to track and extract features associated with mouth shape and appearance.

Landmark-Based Features: Landmark-based features pertain to the identification of specific facial landmarks, either through manual or automated methods. These landmarks may include the corners of the mouth, the tip of the nose, and the outer corners of the eyes. The relationships, including distances and angles, between these identified landmarks serve as valuable features.

Deep Learning-Based Feature Extraction:

The advent of deep learning methodologies, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has significantly transformed the process of feature extraction in Visual Speech Recognition (VSR) and Visual Speech Recognition (VSRec).



Convolutional Neural Networks (CNNs):

CNNs excel in deriving hierarchical feature representations from visual inputs. In the context of VSR and VSRec, CNNs are predominantly employed to extract spatial features from the mouth area.

2D CNNs: These networks are designed to analyze video frames or still images of the mouth region directly. They utilize convolutional filters to identify spatial features, followed by pooling layers that facilitate dimensionality reduction and fully connected layers that assist in classification.

3D CNNs: 3D CNNs are capable of processing video sequences directly, enabling the capture of both spatial and temporal information. They employ 3D convolutional filters that function across spatial dimensions as well as the temporal dimension, thereby effectively capturing the dynamics associated with mouth movements.

3.2 Recurrent Neural Networks (RNNs):

RNNs are specifically engineered to handle sequential data, making them particularly effective in capturing the temporal characteristics of speech. In the context of Visual Speech Recognition (VSR), RNNs are frequently employed to model the sequence of visual features derived from the mouth area.

Long Short-Term Memory (LSTM): LSTM networks represent a specialized form of RNN that adeptly manages long-range dependencies within sequential data. They are extensively utilized in VSR and Visual Speech Recognition (VSRec) to represent the temporal dynamics associated with mouth movements.

Gated Recurrent Units (GRUs): GRUs serve as a more streamlined alternative to LSTMs, achieving comparable performance while necessitating fewer parameters. They are also prevalent in applications of VSR and VSRec.

Hybrid CNN-RNN Architectures: These architectures leverage the advantages of both Convolutional Neural Networks (CNNs) and RNNs. CNNs are tasked with extracting spatial features from each frame of the video sequence, while RNNs are responsible for modeling the temporal dynamics of these features. This integration has yielded promising outcomes in both VSR and VSRec.

3.3 Transfer Learning and Pre-trained Models:

Transfer learning entails the use of pre-trained models developed on extensive datasets (such as ImageNet) and subsequently fine-tuning them for the specific applications of VSR or VSRec. This approach can lead to substantial enhancements in performance, particularly when working with limited training data. Commonly utilized pre-trained models include ResNet, VGGNet, and MobileNet.

4. Feature Fusion and Temporal Modeling:

To improve the efficacy of Visual Speech Recognition (VSR) and Visual Speech Recognition (VSRec) systems, the application of feature fusion and temporal modeling techniques is frequently utilized.

4.1 Feature Fusion:

Early Fusion: This method entails the concatenation of various feature types, such as appearance-based and geometric-based features, prior to their input into a classifier.
Late Fusion: This approach consists of training distinct classifiers on different feature types and subsequently merging their outputs through techniques like weighted averaging or majority voting.
Intermediate Fusion: This technique involves the integration of features at intermediate layers within a deep neural network.

4.2 Temporal Modeling:

Hidden Markov Models (HMMs): HMMs serve as statistical models that can effectively represent the temporal sequence of visemes or the dynamic characteristics of speakers over time.Connectionist Temporal Classification (CTC): CTC functions as a loss function that facilitates the training of Recurrent Neural Networks (RNNs) on unsegmented sequences of visual speech data.
Attention Mechanisms: These mechanisms enable the network to concentrate on the most pertinent segments of the input sequence.

5. Comparison and Evaluation:

The various feature extraction methodologies demonstrate distinct advantages and limitations.

Appearance-based features: These are straightforward to implement but are vulnerable to variations in illumination and pose.

Geometric-based features: These offer greater resilience to illumination changes but necessitate precise landmark detection.

Deep learning-based features: These can capture intricate feature representations but demand substantial training data and are computationally intensive.

The assessment of VSR and VSRec system performance typically employs metrics such as Word Error Rate (WER) for VSR and Equal Error Rate (EER) for VSRec. Standard datasets, including the AVLetters and AVDigits datasets for isolated word recognition, as well as the GRID corpus for continuous speech, are commonly utilized.

6. Applications:

The applications of Visual Speech Recognition (VSR) and Visual Speech Recognition for biometric purposes (VSRec) are varied and expanding:

Assistive Technologies: VSR can aid individuals with hearing impairments in comprehending speech, especially in environments with significant background noise.

Security Systems: VSRec can facilitate biometric identification and verification, thereby enhancing security measures.

Human-Computer Interaction: VSR can support hands-free operation of devices and applications.

Robotics: Robots can utilize VSR to interpret human commands and engage with people in a more intuitive manner.

Silent Speech Interfaces: VSR can be employed to create silent speech interfaces for those with speech disabilities.

7. Conclusion and Future Directions:

Feature extraction plays a vital role in the functionality of VSR and VSRec systems. While conventional appearance-based and geometric-based features have been extensively utilized, recent advancements in deep learning techniques have demonstrated considerable progress. Future research avenues may include:

Developing more resilient and efficient deep learning architectures: This encompasses the exploration of innovative Convolutional Neural Network (CNN) architectures, Recurrent Neural Network (RNN) architectures, and attention mechanisms.

Tackling the issues of speaker variability and background noise: This involves creating methods for domain adaptation and noise mitigation.
Integrating visual and auditory data: This includes the development of multi-modal models capable of effectively merging visual and auditory signals.

Investigating the application of 3D facial models: This entails utilizing 3D facial models to derive more precise and reliable features.

Creating privacy-preserving VSR and VSRec systems: This involves examining techniques such as federated learning to train models without directly accessing sensitive user information.

By persistently enhancing feature extraction methodologies, researchers can facilitate the development of more accurate, robust, and practical VSR and VSRec systems. This progress will significantly influence a range of applications, from assistive technologies to security systems and human-computer interaction, through the integration of advanced feature extraction with sophisticated temporal analysis.

REFERENCES

[1] H. Ertan Çetingül "Discriminative Analysis of Lip Motion Features for Speaker Identification and speech-Reading " Student Member, IEEE, Yücel Yemez, Member, IEEE, Engin Erzin, Member, IEEE, andA. Murat Tekalp, Fellow, IEEE .

[2] Salma Pathan, Archana Ghotkar "Recognition of spoken English phrases using visual features extraction and classification" Department of Computer Engineering Pune Institute of Computer Technology Pune, India

[3] H. Nock and S. Young, “Loosely-Coupled HMMs for ASR,” in Proc. ICSLP, 2000.

[4] K. Saenko, M. Siracusa, K.Wilson, K. Livescu, J. Glass, and T. Darrell, “Visual Speech Recognition with Loosely Synchronized Feature Streams,” in Proc. International Conference on Computer Vision, 2006.

[5] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, “DBN based multi-stream models for audio-visual speech recognition,” in Proc. ICASSP, 2004.

[6] J. C. Wojdel and L. J. M. Rothkrantz, “Using Aerial and Geometric Features in Automatic Lipreading”, in Proceedings Eurospeech 2001, (Scandinavia), September 2001. 2

[7] Yao WenJuan, Liang YaLing, Du MingHui "A Real-time Lip Localization and Tacking for Lip Reading",2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)