Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-024-02138-9

ORIGINAL RESEARCH

Spatio-temporal Weber Gradient Directional feature for visual and audio-visual phrase recognition systems

Salam Nandakishor[1] · **Debadatta Pati**[1]

Received: 26 April 2024 / Accepted: 3 August 2024
© Bharati Vidyapeeth's Institute of Computer Applications and Management 2024

Abstract Visual phrase recognition needs lip movement related visual features, while audio-visual phrase recognition requires both acoustic and visual features. In this work, we propose a novel visual feature; Spatio-temporal Weber Gradient Directional (SWGD) to effectively represent the micro-patterns of lip movements. The proposed visual feature is obtained by using micro-texture information; local differential excitation, gradient orientation, and gradient directional information. Experiments are conducted using standard OuluVS database. Polynomial kernel based support vector machine (SVM) classifier is employed, as it provides relatively better performance. The SWGD extracted from $2 \times 5 \times 3$ video block size provides higher performance of 73.9%. Additionally, we explore twelve distinct local descriptors commonly employed in face recognition and utilize them for the first time in a comparative study of phrase recognition. SWGD performs better than these twelve distinct features but has higher dimension of 4320. By reducing the dimension to 100 using the soft locality preserving map (SLPM), performance improved from 73.9 to 81.3%. The dimensionally reduced SWGD ($\mathrm{SWGD}_{\mathrm{SLPM}}$) outperforms other state-of-the-art visual features mentioned in this paper. This shows the benefit of the salient micro-texture information considered in the proposed feature but neglected in state-of-the-art features. We observe that the $\mathrm{SWGD}_{\mathrm{SLPM}}$ feature has high discriminative ability to represent distinct lip movement patterns for different phrases. Mel-frequency cepstral coefficient (MFCC) based audio phrase recognizer performance degrades as the signal-to-noise level decreases. Including the $\mathrm{SWGD}_{\mathrm{SLPM}}$ visual feature and Glottal MFCC (GMFCC) excitation source feature improves performance by 3.6%, reflecting noise robustness.

Keywords SWGD · $\mathrm{SWGD}_{\mathrm{SLPM}}$ · SLPM · MFCC · GMFCC

1 Introduction

Visual phrase recognition (VPR) is the task of recognizing spoken phrases based on the information of lip movement patterns. Audio-visual phrase recognition (AVPR) is also the task of recognizing spoken phrases, but based on acoustic information and corresponding lip movement information. The VPR and AVPR tasks performed by machines are very useful for many applications, more importantly for hearing impaired people [1] and biometric authentication applications. Moreover, lip movement related visual features are not affected by acoustic noise, and are therefore more robust against unfamiliar conditions. Such visual features are preferably used as supplementary information of acoustic features for developing robust AVPR system.

The lip movement related visual features can be divided into four groups; (1) Geometric visual features [2, 3], (2) Motion visual features [4, 5], (3) Hybrid features [6, 7] and (4) Appearance based visual features [8, 9]. Geometric visual features such as the mouth's height and width [3] require a reliable lip contour tracking method and an accurate face detection algorithm [3]. This task is very difficult when the facial image includes beards and mustaches [10]. The optical flow technique can be used to estimate the motion visual features of the mouth region images [5], but

✉ Salam Nandakishor
salamnandu@gmail.com

Debadatta Pati
debapati2003@yahoo.com

[1] Department of ECE, NIT Nagaland, Chumukedima, Dimapur 797103, Nagaland, India

Published online: 12 August 2024

this approach is sensitive to the speaker's facial orientation and motion. In [7], the authors use lip movements related geometric and motion visual features together. They referred to it as hybrid feature. This approach is effectively able to recognize the English words, but with the support of the corresponding acoustic information. A lip contour tracking algorithm is not needed for the appearance-based visual feature extraction approach. Hence, this method of feature extraction is straightforward.

The appearance based visual features can be classified into two types: (1) global appearance based feature and (2) local appearance based feature. The global appearance based features are estimated directly from the entire region of the image using the discrete cosine transform (DCT) [8] and discrete wavelet transform (DWT) [9], thereby reflecting the global information. Instead of considering the entire region of the image, the local appearance information is extracted from the small regions or patches of the image to extract the micro-patterns of the image. The local appearance based features are also known as local descriptors. An example of such local descriptor is "Local Binary Pattern" (LBP) which describes the gray intensity variation of image patches. This feature represents the appearance or spatial information extracted in XY plane of the image, but not the temporal information. In [11], the authors acquire both spatial and temporal information by extracting the LBP in three orthogonal planes (TOP); XY, XT and YT planes of mouth region images. They termed this feature as spatio-temporal LBP feature or LBP-TOP feature. The histogram patterns of this feature are generated based on the gray intensity difference between the center and neighborhood pixels of mouth region image patches.

All of the above mentioned feature extraction approaches didn't include differential excitation, gradient orientation, and gradient directional information, which provide salient micro-texture information about the image. The differential excitation is the ratio of the gray intensity difference between the central pixel and its neighborhood pixels to the central pixel gray value of the image patch. The gradient orientation is the angle between the vertical and horizontal intensity differences of neighborhood pixels. The last component; gradient directional information is defined by neighborhood pixels gray intensity differences at four different directions; vertical, horizontal, and two slant directions of the image patch. The histogram patterns of the first two components are generated by using the Weber Local Descriptor (WLD) [12] approach, while the histogram of the last component is generated by using the Gradient Direction Pattern (GDP2) [16, 18] approach. We conjecture that these histogram patterns could represent the micro-patterns of lip movements. Hence, the histogram patterns are derived from three planes of mouth region images and concatenated to obtain the proposed visual feature; Spatio-temporal Weber Gradient Directional (SWGD).

The LBP-TOP local descriptor is already explored for visual speech recognition tasks [11, 14]. However, there are many other local descriptors that are commonly used for face recognition tasks [15]. The commonly used twelve unique local descriptors include Binary Pattern of Phase Congruency (BPPC) [17], GDP2 [16, 18], Gradient Local Ternary Pattern (GLTP) [19], Local Directional Pattern (LDP) [20], Local Gradient Increasing Pattern (LGIP) [21], Local Gradient Pattern [22], Local Monotonic Pattern [23], Local Phase Quantization (LPQ) [24, 25], Local Transitional Pattern (LTP) [26], Median Ternary Pattern (MTP) [27], Pyramid of Histogram of Oriented Gradients (PHOG) [28] and WLD [12]. In this work, we extracted these local descriptors in three planes of mouth region images to obtain both the spatial and temporal information of lip movement patterns. The potential of the proposed SWGD visual feature is demonstrated by experiments and a comparative study with twelve different local spatio-temporal features. The dimension of the SWGD feature is reduced by using the soft locality preserving map (SLPM) [29]. This improves performance by increasing the feature's discriminatory ability to classify different phrases. The dimensionally reduced SWGD feature is denoted by $\text{SWGD}_{\text{SLPM}}$.

The performance of an audio-based speech recognizer is degraded in noisy environments due to the distortion of audio speech signals by acoustic noises [30, 31]. However, the visual feature is not affected by acoustic noises [32, 33]. The main aim of this work is to develop a robust phrase recognition system. So, we use both $\text{SWGD}_{\text{SLPM}}$ visual and speech related audio information together for the development of a robust AVPR system. The speech related audio information is prominently represented by the characteristics of the time varying vocal-tract system and time-varying excitation source. In an audio-visual speech recognition system, mostly the Mel-frequency cepstral coefficient (MFCC) [34] feature is used as audio feature to represent the vocal-tract characteristics, and the glottal flow derivative (GFD) [35] wave related features are used to represent the excitation source characteristics. In our previous work [34], we used MFCC and glottal MFCC (GMFCC) together to improve the performance of the audio-visual speech recognition system. This work motivates us to explore MFCC and GMFCC features together as the representation of the audio information for a robust audio-visual phrase recognition task. In the case of excitation source information representation, the estimation approach of the GFD signal is very important for using excitation source information as a supplementary evidence in audio-visual speech recognition. There are many methods available in the literature for GFD estimation. The most efficient methods include iterative adaptive inverse filtering (IAIF) [35, 36], Dynamic

Programming Phase Slope Algorithm (DYPSA) [37], zero-frequency resonator (ZFR) [38], speech event detection using the residual excitation and a mean based signal (SEDREAMS) [39], yet another glottal closure instants (GCI) algorithm (YAGA) [40] and dynamic plosion index (DPI) algorithms [41]. With all these approaches, we prefer to make a comparative study and select the best possible method of GFD estimation to extract the GMFCC excitation source feature for phrase recognition task.

The main contributions of the work presented in this paper are: (1) Proposed spatio-temporal visual features; SWGD and $\text{SWGD}_{\text{SLPM}}$ for visual and audio-visual phrase recognition systems, (2) Explored the twelve different local descriptors commonly used in face recognition, and applying them to visual phrase recognition for a comparative study, (3) Finding the suitable GFD estimation method to extract the GMFCC feature for AVPR system, (4) Analyzing the advantages of using the $\text{SWGD}_{\text{SLPM}}$ visual feature in together with MFCC and GMFCC audio features for audio-visual phrase recognition in different noisy conditions.

The rest of the paper is organized as follows: The literature survey of visual features used in lip reading is given in Sect. 2. The proposed visual feature is discussed in Sect. 3. The description of the database used for experimental analyses is provided in Sect. 4. Experimental results are discussed in Sect. 5. The summary and future scope are reported in Sect. 6.

2 Literature survey

In this section, we report the relevant works and compare the visual features that were employed in visual speech recognition tasks. Additionally, we carefully examine each feature to ascertain its benefits and drawbacks.

In this work [13], authors used three different approaches; Active Shape Model (ASM), Active Appearance Model (AAM) and Multiscale spatial analysis (MSA) to represent lip movement patterns. ASM utilizes statistical models constructed from annotated training images to represent the shape variability of lips. Unlike the ASM, which focuses primarily on shape, AAM considers both shape and appearance simultaneously. The third approach employs a nonlinear scale-space decomposition sieve algorithm to transform the images into a scale-space domain. The temporal information of the lip movements was not taken into account by any of these methods. The experimental studies were carried out with AVletters database [13].

In another work [11], authors proposed a visual that includes both spatial and temporal information of the lip movements. The binary codes or vectors of LBP obtained from XY planes gave the spatial information, whereas the temporal information such as horizontal and vertical motion of the lip movements, was described by feature vectors extracted from XT and YT planes of images. The distributions or histograms of these feature vectors in three planes were concatenated to obtain LBP-TOP features. They compared the performance of the proposed feature with the shape, motion, and global appearance visual features.

Authors [42] proposed a visual feature by combining the planar and stereo information of the global appearance visual feature (DCT) and local appearance visual feature (LBP-TOP) together. Directly concatenating these features would produce a very high dimensional feature. Hence, they reduced the dimension of DCT and LBP-TOP features by using Linear Discriminant Analysis (LDA) and minimal-Redundancy-Maximal-Relevance (mRMR) respectively and then concatenated them. The dimension of this concatenated feature is further reduced by using LDA approach. They termed these final feature vectors as Cascade Hybrid Appearance Visual Feature (CHAVF). This visual feature was employed for connected digit and isolated phrase recognition.

In this work [14], the authors employed the Phase based Eulerian video magnification (EVM) method to acquire the subtle patterns of lip movements by magnifying the input video. First, the desired frequencies of pyramid levels were amplified and passed through the temporal filter. A magnification factor was applied to the temporal filter's output, and produced magnified video. Then, the compact representation of lip movement patterns was obtained from the magnified video using LBP-TOP feature extraction process. We denoted this feature as EVM + LBP-TOP. The support vector machine (SVM) classifier was employed to recognize the phrases.

The literature review is summarized in Table 1. The ASM feature represents the geometric representation or shape of the lip contours. This approach requires manual annotation of the lip contours, which takes a lot of time. We observe that local spatiotemporal visual feature; LBP-TOP outperforms AAM, ASM, MSA, optical flow and DCT based features for visual speech recognition tasks. This is because the local spatio-temporal visual feature acquires both local spatial and temporal information that effectively represent the lip movement patterns. The performance of LBP-TOP feature was improved by magnifying the video using EVM algorithm. However, the video magnification algorithm is time consuming process. So it will be difficult to employ in real-time applications. For small database such as AVletters database, the SVM classifier outperforms the HMM modeling technique for LBP-TOP feature. It means the choice of classifier has an impact on the performance of the visual speech recognition system. Since, OuluVS database [11] is a small database, so for developing phrase recognition system, we have chosen to use SVM classifier.

Table 1 Summary of the literature survey on different types of visual features

Sl. no	Feature	Feature type	Classifier	Database	Accuracy%
1	ASM[13]	Shape	HMM	AVletters	26.90
2	AAM[13]	Hybird	HMM	AVletters	41.90
3	MSA[13]	Global appearance	HMM	AVletters	44.60
4	LBP-TOP [11]	Local Spatio-temporal	HMM	AVletters	57.30
5	Optical Flow [11]	Motion	SVM	AVletters	32.31
6	DCT [11]	Global appearance	SVM	AVletters	51.15
7	LBP-TOP [11]	Local Spatio-temporal	SVM	AVletters	58.85
8	LBP-TOP [11]	Local Spatio-temporal	SVM	OuluVS	64.20
9	DCT + LDA [42]	Global appearance with LDA	SVM	OuluVS	64.93
9	CHAVF [42]	Hybrid	HMM	OuluVS	68.90
10	EVM + LBP-TOP [14]	EVM + Local Spatio-temporal	SVM	OuluVS	70.00

The gradient orientation, differential excitation, and gradient directional information that provide important micro-texture information about the mouth portion of images were not included in the visual feature extraction methods discussed in the literature review. Hence, we propose visual feature; SWGD to represent the lip movement patterns efficiently. The dimension of the proposed visual feature is high. So, we reduce the feature dimension by using SLPM algorithm.

3 Proposed methodology

In this section, we discuss the processing steps of proposed visual feature extraction. First the facial portion is detected and then cropped the mouth portion automatically. The differential excitation, gradient orientation, and gradient directional information are estimated from mouth region images. The histogram of WLD are generated by using the differential excitation and gradient orientation information, whereas the histogram of GDP2 is generated by using the gradient directional information. These histograms are obtained in XY, XT and YT planes to acquire the both spatial and temporal information. Then, these histograms are concatenated together to obtain the SWGD visual feature.

The facial portion detected by Viola-Jones face detection algorithm is further processed to crop the mouth region. In order to extract the spatio-temporal features of lip movements, it is very important to localize the mouth region accurately. The detected facial images are divided into blocks for finding the "Region of Interest" or "Mouth Region". By conducting the empirical analysis, we decided to create 10 horizontal blocks and 11 vertical blocks of the detected facial image. These 10 horizontal and 11 vertical blocks are obtained by dividing the number of rows of image by 10 and number of columns by 11. The common portion between the last 3 horizontal blocks and vertical blocks (from 3^{rd} to 8^{th} vertical blocks) is considered as "Region of Interest", where the mouth portion exist well. The cropped mouth region frames extracted from the input video is not equal in size. It means the number of rows and columns of the mouth region images are not same. However, to extract the spatio-temporal visual feature, frames present in each video should have an equal number of rows and columns. Therefore, we resized the mouth region image frames to maintain a uniform frame size for each video.

The differential excitation, gradient orientation, and gradient directional information are estimated from mouth region images or frames to represent the distinctive patterns of lip movements. While uttering speech, different patterns of lip movements are generated. These patterns can be represented by histogram of feature vectors generated from the mouth region images. The differential excitation and gradient orientation feature vectors are generated by using WLD, whereas the gradient directional feature vector is obtained by using GDP2. The histogram patterns of feature vectors are concatenated to obtain the proposed visual feature; SWGD.

The differential excitation (Ψ) is calculated by using the following equation.

$$\Psi = arctan\left(\sum_{x=0}^{S-1}\left(\frac{(I_x) - (I_c)}{I_c}\right)\right) \tag{1}$$

where S is the total number of neighbor pixels. The intensity values of neighborhood pixels and central pixels are denoted by I_x and I_c respectively. Every central pixel has eight neighborhood pixels, therefore the value of S is 8.

The numerator of Eq. 1 is the sum of the differences of intensity value of neighboring pixels against its central pixel, whereas the denominator represents the intensity value of central pixel. For simplicity, the values of Ψ are quantized into N dominant differential excitation by using Eq. 2. By empirical analysis, the value of N is set to 8 for our proposed visual feature.

$$\Psi_n = floor\left(\frac{\Psi + \pi/2}{\pi/N}\right), \quad n = 0, 1, 2, ..., N - 1 \tag{2}$$

The gradient orientation (Θ) is defined by the equation below.

$$\Theta = arctan\left(\frac{I_5 - I_1}{I_7 - I_3}\right) = arctan\left(\frac{I_V}{I_H}\right) \tag{3}$$

where I_5 and I_1 are the intensity values of lower and upper neighbors pixels of central pixel I_c whereas I_7 and I_3 are the intensity values of left and right neighbors pixels of I_c.

The range of Θ is within {-π/2, π/2}. To obtain more information about the gradient direction, the range of Θ is increased by mapping to $\Theta' \in [0, 2\pi]$. This mapping is done according to the values of (I_V) and (I_H).

$$\Theta'(x,y) = \begin{cases} \Theta(x,y), & I_V(x,y) > 0, I_H(x,y) > 0 \\ \Theta(x,y) + \pi, & I_V(x,y) < 0, I_H(x,y) > 0 \\ \Theta(x,y) + \pi, & I_V(x,y) < 0, I_H(x,y) < 0 \\ \Theta(x,y) + 2\pi, & I_V(x,y) > 0, I_H(x,y) < 0 \end{cases} \tag{4}$$

The values of Θ' are quantized to T dominant gradient orientation by using Eq. 5. In this work, the value of T is set to 4 because the proposed feature produced good performance at this value.

$$\phi_t = floor(\frac{\Theta'}{2\pi/T}), \quad t = 0, 1, 2, ..., T - 1 \tag{5}$$

By using the quantized differential excitation (Ψ_n) and quantized gradient orientation (ϕ_t), the 2D histogram of {WLD(Ψ_n, ϕ_t)} is generated. Then the 2 dimensional WLD histogram is transformed to row feature vectors. The length of WLD feature vectors is $N \times T$.

It is conjectured that the WLD feature may be suitable for visual and audio-visual phrase recognition applications. Therefore, we explored this local descriptor for our task objective. The gradient orientation component of WLD considers only vertical and horizontal gradient directions information estimated by {I_1 and I_5} and {I_3 and I_7} pixels. The slanting gradient directions between {I_2 and I_6} and {I_4 and I_8} pixels are not considered. This can be observed from Eq. 3 of gradient orientation calculation of the WLD feature.

In order to include the slanting gradient direction, we employ another type of local descriptor; GDP2 which describes the gradient direction information of four directions; vertical, horizontal and two slant directions. In GDP2 feature extraction, the gray intensity values of pixels at eight different directions (East, North East, North, North West,West, South West, South and South East) are considered. Then the differences between intensity values of pixels at vertical, horizontal, and two slant directions are calculated. The Gradient Direction Pattern feature is generated from sum of gray intensity differences of four directions [16].

Instead of calculating the WLD and GDP2 local descriptors only in the spatial domain (XY plane), we also extracted these features from XT and YT planes to include the temporal information. The WLD and GDP2 features generated from XY, XT and YT planes are concatenated to produce the proposed visual feature; SWGD as shown in Fig. 1.

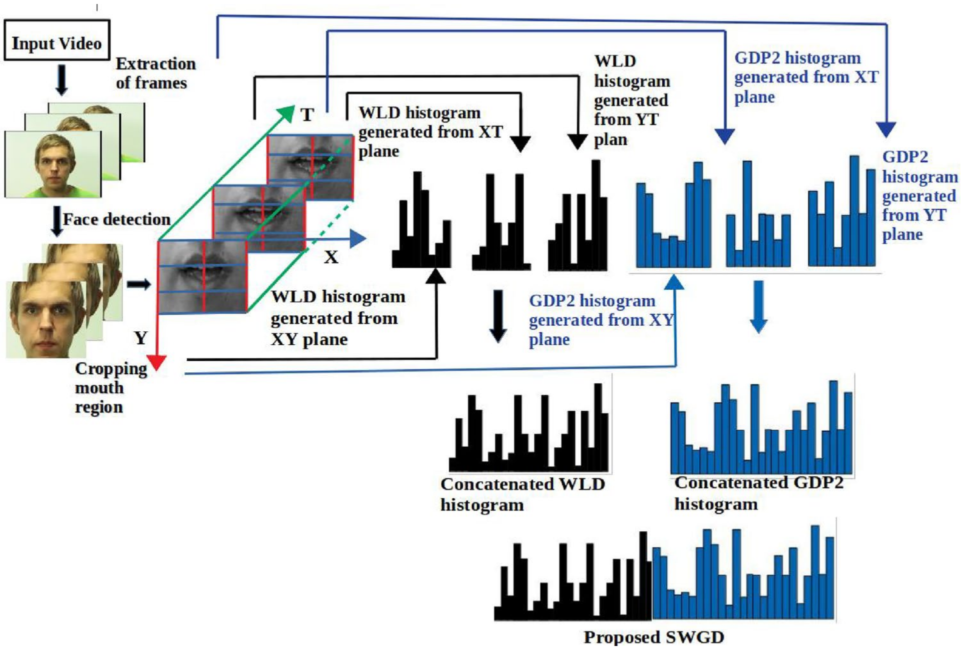
Input Video
Extraction of frames
Face detection
Cropping mouth region
T
X
Y
WLD histogram generated from XT plane
WLD histogram generated from YT plan
GDP2 histogram generated from XT plane
GDP2 histogram generated from YT plane
WLD histogram generated from XY plane
GDP2 histogram generated from XY plane
Concatenated WLD histogram
Concatenated GDP2 histogram
Proposed SWGD

Fig. 1 Methodology of the proposed visual feature extraction algorithm

4 Database description

The experimental studies are carried out with the OuluVS audio-visual database [11]. This database was recorded by 20 speakers (17 males and 3 females). Out of 20 speakers, 9 speakers wear glasses. Each speaker read 10 phrases repeated up to 5 times to make suitable for visual and audio-visual phrase recognition experiments. The details of the phrases with assigned labels are given in Table 2. The speakers belong to four different countries with different speaking speeds and accents. This database was recorded inside the controlled environment room. The distance between the speaker and the camera is maintained at 160 cm. The frame rate of the video data is 25 frames per second (fps) and the image resolution is 720 × 576 pixels.

5 Experimental result and discussion

First, we did the experiments for different twelve local descriptor based visual features, and SWGD for ten phrases recognition. The performance of SWGD is also analyzed by using three different block sizes ($2 \times 5 \times 3$, $2 \times 3 \times 2$ and $2 \times 2 \times 2$) to select the best video block size. The dimension of the SWGD feature is significantly higher. Hence, we reduced the feature dimension by using the SLPM method and analyzed the feature vectors distribution by using t-SNE (t-distributed stochastic neighbor embedding) plots.

As mentioned in the earlier section, there are various methods for GFD estimation. Therefore, we conducted a comparative analysis to select the best GFD estimation method for extracting GMFCC. Then, this glottal excitation source feature; GMFCC is concatenated together with vocal tract feature; MFCC and visual feature; $\text{SWGD}_{\text{SLPM}}$ to develop an audio-visual phrase recognizer.

5.1 Visual phrase recognition experiments

We used SVM classifiers for training and testing the visual and audio-visual phrase recognition systems. To obtain the optimum accuracy of the SVM classifier, multiple train and test data sets are created using the “Leave-One-Out” cross-validation approach. Each test data set considers only one utterance for each phrase of each speaker, and the remaining utterances for each phrase of each speaker are considered in the training data set. This step is repeated for all the utterances. The final accuracy of the system is calculated by averaging the individual scores. This cross-validation method is a time consuming approach. However, it is suitable for OuluVS database because the number of utterances or samples for each phrase of each speaker present in the database is not large scale.

Table 2 P1 to P10 labels represent ten phrases respectively

Label	Phrase	Lable	Phrase
P1	Excuse me	P6	See you
P2	Good bye	P7	I am sorry
P3	Hello	P8	Thank you
P4	How are you	P9	Have a good time
P5	Nice to meet you	P10	You are welcome

The local descriptors mentioned in Sect. 1 are successfully employed in image pattern recognition applications like face recognition. These types of local descriptors each have their own merits and demerits. For example, a local descriptor like LPQ is computationally simple, but they are sensitive to noise and illumination variation. Similarly, Gabor-based local descriptor like BPPC is insensitive to illumination variation. However, it faces the problem of high computational requirements and high feature dimensionality. Therefore, finding a robust and discriminative local descriptor is still an interesting research area for image pattern representation and classification.

In this work, we extracted the spatio-temporal information of twelve different local descriptors and evaluated the performance for visual phrase recognition experiments. The local descriptors are extensively used for face recognition. However, only a few local spatio-temporal descriptors such as LBP-TOP feature have been used for the visual speech recognition system. Therefore, we explore other types of local spatio-temporal descriptors particularly for visual phrase recognition applications. They are GDP2-TOP, GLTP-TOP, BPPC-TOP, LDP-TOP, LGIP-TOP, LGP-TOP, LMP-TOP, LPQ-TOP, LTP-TOP, MTP-TOP, PHOG-TOP and WLD-TOP. We compared the performance of these features with the proposed SWGD visual feature by using SVM classifier.

The performance of the SVM classifier depends on the type of kernel function. So, we compare the performance of visual phrase recognition by using three different kernel functions; Polynomial, Linear and Radial basis function (RBF). The experimental results are given in Table 3. From the results, it is evident that for all local spatio-temporal features, SVM with polynomial kernel function performs better than linear and RBF kernel functions. Further, our proposed SWGD visual feature provides higher accuracy than all twelve different local spatio-temporal descriptors. It is because the proposed visual feature considers important micro-texture information such differential excitation, gradient orientation and gradient directional information together for representing the patterns of lip movements.

From Table 4, we can observe that the proposed visual feature extracted from the block size of $2 \times 5 \times 3$ in XY, XT and YT planes produced the best possible result. Therefore, this video block size is considered for all the experimental

Table 3 Accuracies (in %) of visual phrase recognition system with different spatio-temporal features

Sl. No	Visual feature	Polynomial	Linear	Radial basis function
1	GDP2-TOP	69.9	62.4	65.4
2	GLTP-TOP	52.5	45.3	44.0
3	BPPC-TOP	64.4	59.7	58.9
4	LDP-TOP	63.8	57.8	55.2
5	LGIP-TOP	60.4	52.8	57.0
6	LGP-TOP	63.7	55.3	58.4
7	LMP-TOP	58.8	53.2	53.1
8	LPQ-TOP	63.5	47.8	50.1
9	LTP-TOP	49.0	43.3	41.8
10	MTP-TOP	66.9	58.3	61.6
11	PHOG-TOP	61.6	55.7	58.7
12	WLD-TOP	72.6	65.7	66.0
13	SWGD	**73.9**	67.0	67.5

Bold represents the best accuracy

Table 4 Performance of VPR system with different block sizes

Sl. No	Block size	Result (%)
1	2 × 5 × 3	**73.90**
2	2 × 3 × 2	72.70
3	2 × 2 × 2	64.30

Bold represents the best accuracy

analyses. Since, the proposed SWGD visual feature is obtained by combining the WLD-TOP and GDP2-TOP features, the dimension of the SWGD feature is the sum of the dimensions of these two features. The feature dimension of WLD-TOP is determined by multiplying the size of the block, the number of planes, and the dimension of the WLD feature. The dimension of the WLD-TOP feature is equal to 2880 {size of block (2 × 5 × 3) × three orthogonal planes (3) × size of WLD histogram (8 × 4)} whereas the GDP2-TOP feature has a dimension of 1,440 {size of block (2 × 5 × 3) × three orthogonal planes (3) × size of GDP2 histogram (16)}. Therefore, the proposed SWGD visual feature has a dimension of 4320 (2880 + 1440). We employed the SLPM approach to reduce the dimension of the SWGD visual feature to 100. We denoted this transformed visual feature by $\text{SWGD}_{\text{SLPM}}$. The SLPM method not only reduces the dimension of the feature but also increases the discriminative ability of the proposed features to classify the phrases.

We used the t-SNE plot to visualize the distribution of visual features of 10 different phrases in two dimensional space. The distributions of phrases plotted by t-SNE plot by using SWGD and $\text{SWGD}_{\text{SLPM}}$ visual features are shown in Fig. 2(a) and (b) respectively. In t-SNE plot, we represent the 10 phrases by using ten different labels. We observed that the distribution of the phrases is very close to each other and difficult to classify. On the other hand, the distribution of phrases using $\text{SWGD}_{\text{SLPM}}$ feature is clearly seen as separable clusters in Fig. 2(b). This shows that the SLPM feature dimensionality reduction approach increased the discriminative ability or separability among the classes. This SLPM approach improved the performance of the proposed SWGD visual feature from 73.9 to 81.30%.

We also compared the performance of our proposed visual feature with six state-of-the-art appearance based visual features that are used in visual phrase recognition. The appearance based state-of-the-art visual features considered in this comparison are LBP-TOP [11], DCT+LDA [42], Sequential Pattern Boosting (SP-Boosting) [43], CHAVF [42],

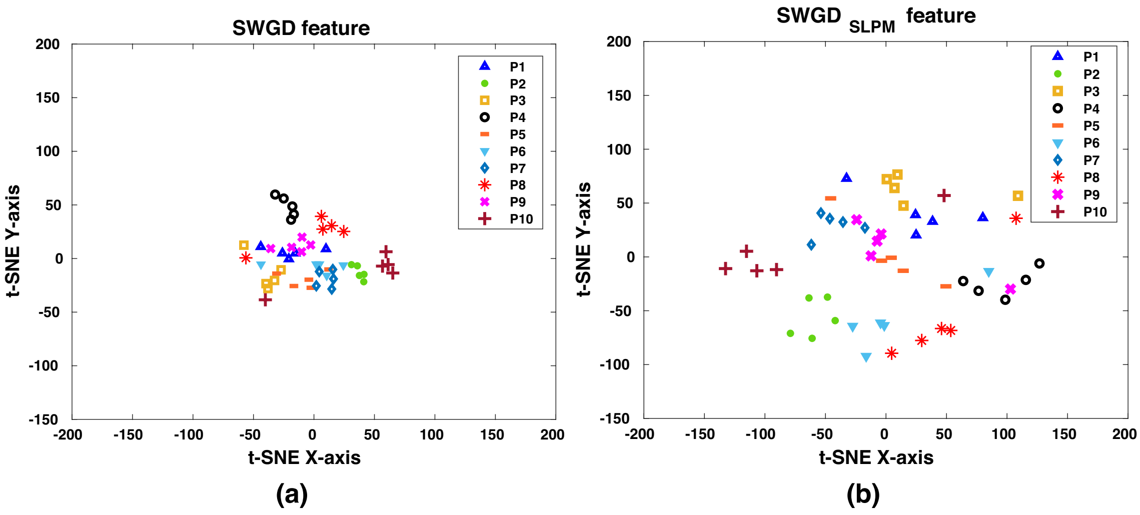

Fig. 2 The t-SNE plot by using **a** SWGD and **b** $\text{SWGD}_{\text{SLPM}}$

Table 5 Performance comparison of proposed visual features with state-of-art visual features on OluVS database for VPR task

Sl. No	Visual feature	Result (%)
1	LBP-TOP [11]	64.20
2	DCT+LDA [42]	64.93
3	SP-Boosting [43]	65.60
4	CHAVF [42]	68.90
5	EVM+LBP-TOP [14]	70.00
6	TSRVFs [44]	70.60
7	SWGD	**73.90**
8	Reduced dimensional SWGD ($SWGD_{SLPM}$)	**81.30**

Bold represents the best accuracy

EVM+LBP-TOP [14] and Transported Square-Root Vector Fields (TSRVFs) [44]. The results of these state-of-the-art features and our proposed feature are given in Table 5. All these reported visual phrase recognition experiments were conducted using the OuluVS database. The LBP-TOP is also a spatio-temporal feature, but it is obtained by thresholding the gray color intensity differences between the center pixel and neighborhood pixels. DCT is the global visual feature, which is extracted from the entire mouth region, and LDA is used to reduce the dimension of the feature. SP-Boosting is the machine learning approach proposed for visual speech recognition in [43] and authors used a visual feature that is related to the intensity differences of the mouth region images. The CHAVF feature proposed in [42] is the combination of local feature LBP-TOP and global feature DCT. In [14], the authors applied the EVM technique to the input videos in order to amplify the subtle information of lip movements. The LBP-TOP feature extracted from this magnified input video is denoted by EVM+LBP-TOP. The LBP-TOP feature with the EVM approach could effectively represent the patterns of lip movements. Nevertheless, real-time lip reading applications are not appropriate for this slow video magnification method. In [44], authors calculated the covariance matrices of pixel location, intensity and their derivatives of the mouth region images and obatined the correlation matrices. They constructed the trajectories of correlation matrices using TSRVF for phrase recognition. All these reported appearance based visual features do not consider the micro-texture information such as differential excitation, gradient orientation, and gradient directional information extracted from the mouth region images. Hence, the proposed visual feature outperformed all those state-of-the-art appearance based visual features and was found suitable for representing the patterns of lip movements for different phrases.

From the experimental results and analysis, we can conclude that the proposed $SWGD_{SLPM}$ is the best possible representation of lip movement patterns for visual phrase recognition. In the following section, we include audio information to develop a relatively more robust audio-visual phase recognition system.

5.2 Audio-visual phrase recognition experiments

The GFD signal can be estimated by using the IAIF, DYPSA, ZFR, SEDREAMS, YAGA, and DPI approaches. The GFD signals estimated by the aforementioned methods can be compared by conducting a discriminative analysis of phrases using GMFCC features. The GFD signal is used as the input signal, and then the MFCC feature extraction procedure is applied to extract the GMFCC feature.

The distribution of the GMFCC feature vectors is plotted in two dimensions using t-SNE. The GMFCC features are extracted from GFD signals that are estimated using the IAIF, DYPSA, ZFR, SEDREAMS, YAGA, and DPI methods. The t-SNE plots are shown in Fig. 3 and the 10 phrases are labeled as P1, P2, P3, P4, P5, P6, P7, P8, P9, and P10. The labels with assigned phrases of OuluVS database are given in Table 2. The distribution of GMFCC features for 10 phrases is clearly seen as separable clusters in Fig. 3(a) whereas the distribution of the phrases shown in Fig. 3(b–f) is very close to each other, making it difficult to classify the phrases. This demonstrates that the IAIF GFD estimation method is more suitable than other methods; DYPSA, ZFR, SEDREAMS, YAGA, and DPI to extract GMFCC feature for phrase recognition. This is due to the fact that while other approaches require the locations of glottal closure instants (GCIs) to be obtained, the IAIF approach does not. It is challenging to accurately estimate the location of GCIs from noisy speech. Therefore, we employed the IAIF method in GFD estimation to extract the GMFCC feature for the AVPR system.

The main objective of the proposed AVPR system is to recognize spoken phrases in noisy conditions. In this experiment, we removed the silence portions at the starting and ending parts of the audio speech files, because they do not carry any important information related to the phrases. We use four additive white Gaussian noise levels having SNR varying from – 6dB, – 3dB, + 3dB, to + 6dB to create noisy speech signals. The experimental results are presented in Table 6. Individually MFCC provides the better performance, but MFCC and GMFCC in together further improves the performance, reflecting the usefulness of using the excitation source based GMFCC feature as a supplementary evidence for phrase recognition in noisy conditions.

As we mentioned earlier, the visual information is not affected by noise, and therefore the performance of the proposed $SWGD_{SLPM}$ visual feature remains unchanged at all noise levels. However, the benefit of using the $SWGD_{SLPM}$ visual feature helps in improving the performance. For example, on an average from extremely low (– 6dB) to high noise level (+ 6dB) the performance improved from 90.10

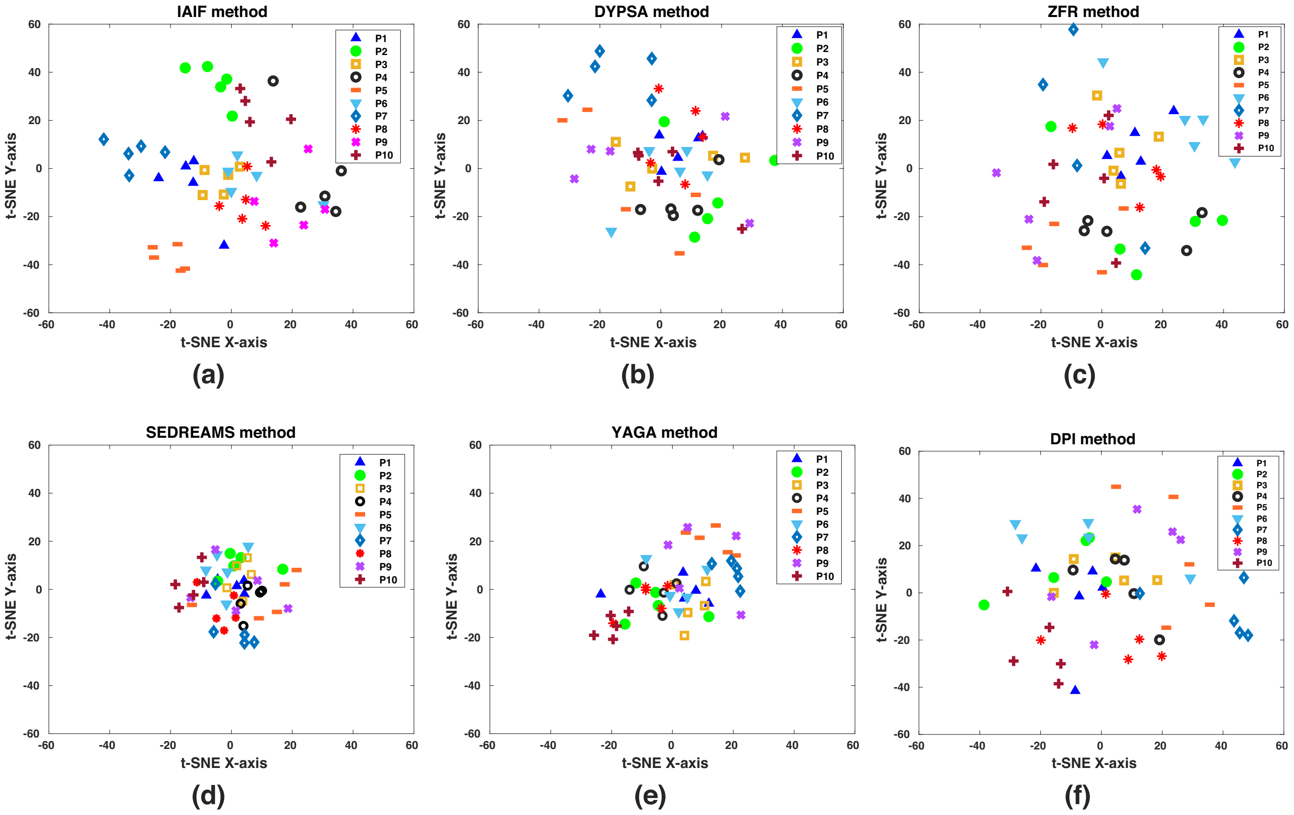

Fig. 3 The t-SNE plot of GMFCC features extracted from GFD signal estimated by using **a** IAIF, **b** DYPSA, **c** ZFR, **d** SEDREAM, **e** YAGA and **f** DPI method. The labels (P1 to P10) represent the ten phrases

Table 6 Performance (in %) of phrase recognition at different SNR levels

Feature	– 6dB	– 3dB	+3dB	+6dB	Avg
MFCC	82.8	87.2	91.9	93.0	88.73
GMFCC	78.5	82.8	89.5	91.4	85.55
MFCC+GMFCC	84.1	88.4	93.4	94.5	90.10
SWGD$_{\mathrm{SLPM}}$	81.3	81.3	81.3	81.3	81.30
MFCC+GMFCC+SWGD$_{SLPM}$	87.1	90.9	94.5	95.0	**91.88**

Bold represents the best accuracy

to 91.88%, reflecting a relative improvement of 2%. It is also observed that at the lowest noise level (– 6 dB), the inclusion of the proposed SWGD$_{\mathrm{SLPM}}$ visual feature significantly helps in improving the performance of the audio-visual phrase recognition.

6 Conclusion and future scope

The proposed SWGD and its dimensionally reduced SWGD$_{\mathrm{SLPM}}$ visual features that give better results than all twelve local spatio-temporal features in the context of phrase recognition tasks. The experiments are made with an internationally standard OuluVS database and a polynomial kernel function based SVM classifier. The experimental results show that the GDP2-TOP and WLD-TOP features are providing better performance of 69.9% and 72.6% respectively, among the twelve local spatio-temporal features. These performances are lower than our proposed SWGD (73.9%) and low dimensional SWGD$_{\mathrm{SLPM}}$ (81.30%) visual features. A comparative study with other state-of-the-art features shows that the TSRVFs feature provides the good performance of 70.6 %, which is less than our proposed SWGD$_{\mathrm{SLPM}}$ visual feature. The reason for getting better performance may be due to the effective representation of micro-texture lip movement patterns using both spatial and temporal information. The use of a reduced dimensional SWGD$_{\mathrm{SLPM}}$ visual feature is another reason to achieve further improvement in the performance.

In our previous work, MFCC, GMFCC, and their combined representation (MFCC+GMFCC) acoustic representations were found to be effective in recognizing confusing phonemes like (‘p’ and ‘b’) and the English letters (‘P’ and ‘B’). Motivated by this work, we use both of these acoustics features; the vocal-tract related feature (MFCC) and the glottal excitation source related feature (GMFCC) for phrase recognition. Here, we first verify the best suitable method of GFD estimation, particularly for extracting the GMFCC feature for phrase recognition task. We observed that the GMFCC feature extracted through the IAIF based GFD estimation method provides better classification of phrases than other GFD estimation methods; DYPSA, ZFR, SEDREAMS, YAGA, and DPI. This is because the IAIF approach does not require the locations of GCIs, but others approaches need to obtain GCIs locations. The accurate estimation of GCIs location from noisy speech is a difficult task. It is concluded that the IAIF method is found suitable for glottal excitation source feature estimation in phrase recognition applications. The experimental results show that the acoustic based phrase recognition system provides better performance than the proposed VSR system. However, with the inclusion of acoustic additive noise, the VPR system performance remains unchanged, but audio based system performance is degraded proportionally with the SNR level of the noise. By including the GMFCC excitation source feature and the proposed $\text{SWGD}_{\text{SLPM}}$ visual feature, the best possible performance of the audio based system has been relatively increased by 3.6%. This shows the robustness of the proposed AVPR system against noise.

The proposed visual features could be used for a continuous audio-visual speech recognition system. The performance of our proposed visual feature could be improved with a larger audio-visual database and deep neural network modeling. This proposed visual feature may be suitable for other speech processing applications, such as audio-visual speaker recognition and language identification.

Author contributions The authors propose a spatio-temporal feature for visual and robust audio-visual phrase recognition systems.

Funding This work is part of the project title “Development of Excitation Source Features Based Spoof Resistant and Robust Audio-Visual Person Identification System” which is funded by Ministry of Electronics and Information Technology, Government of India. Award No: 12(5)/2015-ESD.

Data availability The OluVS database is not an open source audio-visual database; the authors do not have the right to make it available.

Code availability Code will made available on reasonable request.

Declaration

Conflict of interest The authors do not have conflict of interest.

References

1. Sinha GR (2017) Indian sign language (ISL) biometrics for hearing and speech impaired persons: review and recommendation. Int J Inf Technol 9:425–430
2. Kaynak M et al (2004) Analysis of Lip Geometric Features for Audio-Visual Speech Recognition. IEEE Trans on Systems, Man, and Cybernetics 34(4):564–570
3. Salam Nandakishor & Debadatta Pati (2020) Extraction of lip contour and geometric lip features for audio-visual phoneme recognizer. IJCSPL 6(1):25–33
4. Tamura S et al (2004) Multi-modal speech recognition using optical-flow analysis for lip images. J Signal Process Syst 36(3):117–124
5. Sharma Usha et al (2019) Visual speech recognition using optical flow and hidden Markov model. Wireless Pers Commun 106:2129–2147
6. Chan MT (2001) Hmm-based audio-visual speech recognition integrating geometric and appearance-based visual features. In: Conference MMSP
7. Nandakishor S, Pati D (2021) Analysis of Lombard effect by using hybrid visual featuresfor ASR. In: Pattern Recognition and Machine Intelligence
8. Xiaopeng Hong, et al. (2006) A PCA Based Visual DCT Feature Extraction Method for Lip-Reading. In: International Conference, IIH-MSP
9. Puviarasan N, Palanivel S (2010) Lip reading of hearing impaired persons using HMM. Expert Syst Appl 38(4):4477–4481
10. Wang SL et al (2007) Robust lip region segmentation for lip images with complex background. Pattern Recogn 40(12):3481–3491
11. Zhao G, Barnard M, Pietikainen M (2009) Lipreading with local spatiotemporal descriptors. IEEE Trans Multim 11(7):1254–1265
12. Chen J et al (2010) WLD: A Robust Local Image Descriptor. IEEE Trans Pattern Anal Mach Intell 32(9):1705–1720
13. Matthews I et al (2002) Extraction of visual features for lipreading. IEEE Trans Pattern Anal Mach Intell 24(2):198–213
14. Nandakishor S, Pati D (2021) Phrase recognition using Improved Lip reading through Phase-Based Eulerian Video Magnification. In NCC
15. Eleyan Alaa (2023) Statistical local descriptors for face recognition: a comprehensive study. Multim Tools Appl 82:32485–32504
16. Turan Cigdem, Lam Kin-Man (2018) Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study. J Vis Commun Image Represent 55:331–341
17. Shojaeilangari S, et al. (2012) Feature extraction through Binary Pattern of Phase Congruency for facial expression recognition. In Conference ICARCV
18. Islam Mohammad Shahidul, Auwatanamongkol Surapong (2013) Gradient direction pattern: a gray-scale invariant uniform local feature representation for facial expression recognition. J Appl Sci 13(6):837–845
19. Ahmed F, Hossain E (2013) Automated facial expression recognition using gradientbased ternary texture patterns. Chin J Eng 2:1–8
20. Jabid T, et al (2010) Local directional pattern - A robust image descriptor for object recognition. In: Int’l Conf. on Advanced Video and Signal Based Surveillance
21. Lubing Z, Han W (2012) Local gradient increasing pattern for facial expression recognition. In: 19^{th} International Conference on Image Processing
22. Islam MS (2014) Local gradient pattern-A novel feature representation for facial expression recognition. J oAI Data Min 2:33–38
23. Mohammad T, Ali ML (2011) Robust facial expression recognition based on local monotonic pattern. In: Int’l Conf. on Computer and Information Technology
24. Ojansivu V, Heikkila J (2008) Blur insensitive texture classification using local phase quantization. In: International conference on image and signal processing

25. Dhall A, et al. (2011) Emotion recognition using PHOG and LPQ features. In: IEEE International Conference on Automatic Face and Gesture Recognition
26. Jabid T, Chae O (2012) Facial expression recognition based on local transitional pattern. Int J Inform 15(5):2007–2018
27. Bashar F, et al. (2014) Robust facial expression recognition based on median ternary pattern. In: Int'l Conf. on Electrical Information and Comm. Technology
28. Bosch A, et al. (2007) Representing shape with a spatial pyramid kernel. In: 6th ACM international conference on Image and video retrieval
29. Turan C, Lam KM, He X (2018) Soft Locality Preserving Map (SLPM) for Facial Expression Recognition
30. Nisa R, Baba AM (2024) A speaker identification-verification approach for noise-corrupted and improved speech using fusion features and a convolutional neural network. International Journal of Information Technology
31. Kumar A, Mittal V (2021) Hindi speech recognition in noisy environment using hybrid technique. Int J Inf Technol 13:483–492
32. Shashidhar R et al (2022) Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int J Inf Technol 14:3425–3436
33. Chelali FZ (2023) Bimodal fusion of visual and speech data for audiovisual speaker recognition in noisy environment. Int J Inf Technol 15:3135–3145
34. Nandakishor Salam, Pati Debadatta (2023) Usefulness of glottal excitation source information for audio-visual speech recognition system. Int J Speech Technol 26:933–945
35. Alku P (1992) Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun 11(23):109–118
36. Alku P, Vilkman E (1996) A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. IEEE Trans Audio Speech Lang Process 48(5):240–254
37. Naylor PA et al (2007) Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans Audio Speech Lang Process 15(1):34–43
38. Murthy KSR, Yegnanarayana B (2008) Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process 16(8):1602–1613
39. Drugman T, Dutoit T (2009) Glottal closure and opening instant detection from speech signals. In: Interspeech
40. Thomas MR et al (2012) Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans Audio Speech Lang Process 20(1):82–91
41. Prathosh A et al (2013) Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Trans Audio Speech Lang Process 21(12):2471–2480
42. Sui Chao et al (2017) A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition. Speech Commun 90:26–38
43. Ong EJ, Bowden R (2011) Learning sequential patterns for lipreading. In: Proceedings of 22^{nd} British Machine Vision Conference
44. Su J, et al. (2014) Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.