Visual Speech Recognition Using LBP Features

Abhilash Jain   Rathna G. N.
abhilashjain@ee.iisc.ernet.in   rathna@ee.iisc.ernet.in

Department of Electrical Engineering
Indian Institute of Science, Banglaore

Problem Definition

Automatic lip reading, or visual speech recognition, deals with the task of extracting relevant speech information from visual cues from a person's face/mouth region. Three major challenges in building such a system are:

- Accurate lip segmentation and modeling
- Feature extraction
- Appropriate classifier design

Introduction

Current state-of-the-art **Audio Speech Recognition** (ASR) systems are *not very robust to ambient noise* and fail in *presence of multiple speakers.*

Visual Speech Recognition (VSR) or automatic lip reading provides complimentary speech information, and has many applications:

- Improve performance of ASR in presence of acoustic noise
- Extract speech information in audio-less scenarios
- Human-computer interface
- Refine speech degraded due to speech impairments and Lombard effect

Work on VSR has been done as early as 1980s. Since then, however, VSR has seen very little improvements in performance.

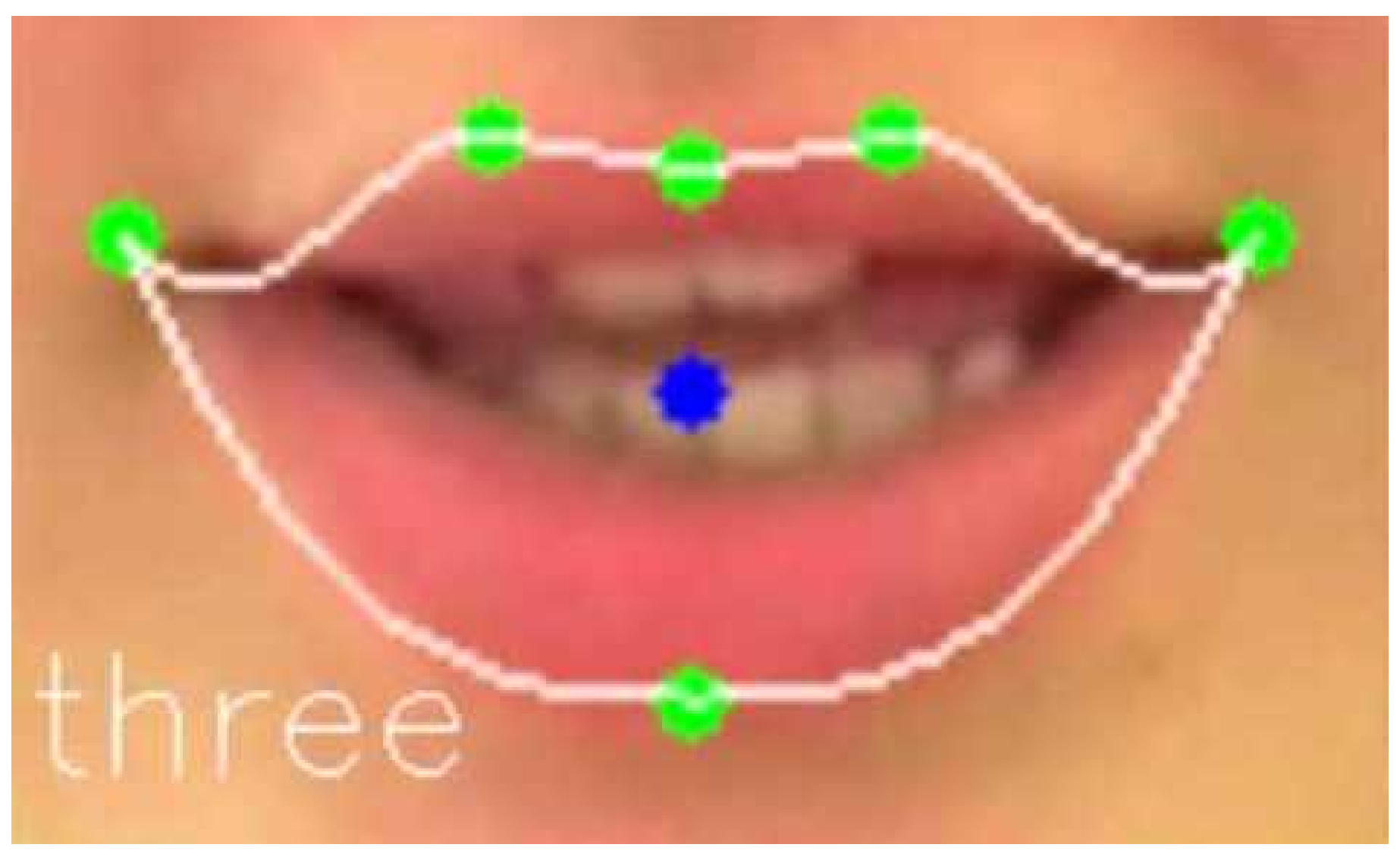


Figure 1: Automatic Lip Reading Model Example

Most VSR systems use the following set of algorithms for feature extraction [1]:

- Active Shape Model (ASM) or Snakes
- Edge-based techniques
- Markov random field based techniques

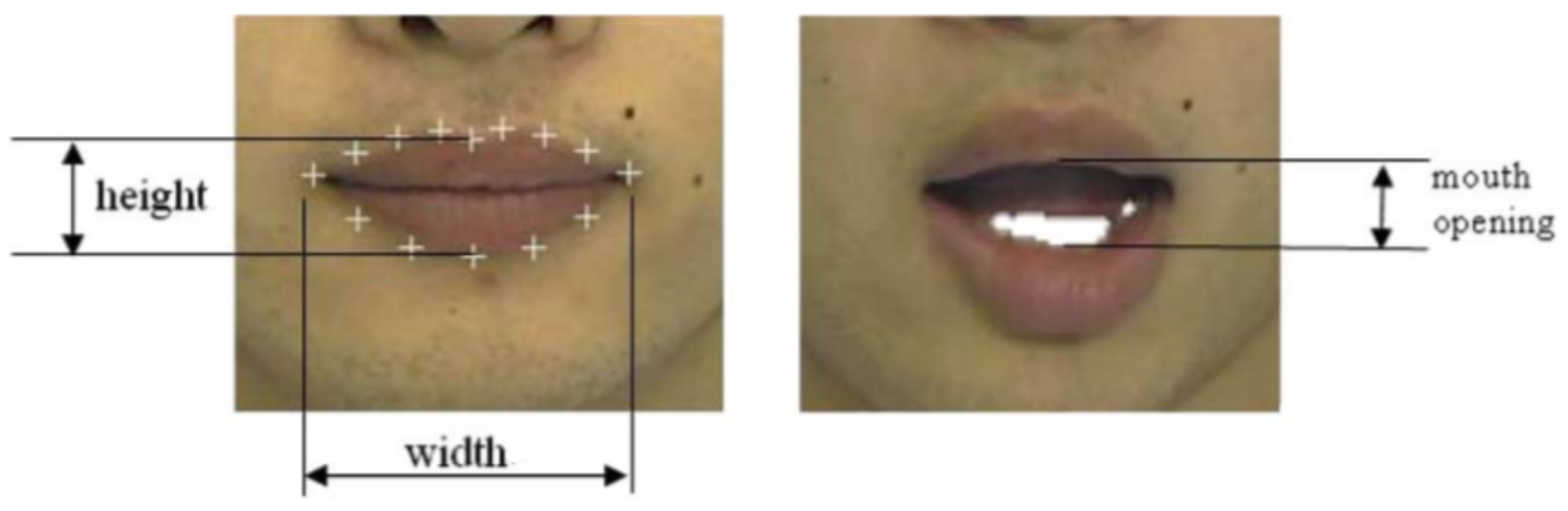


Figure 2: Typical VSR Features

We propose a new feature extractor and classifier which is **not based on lip modeling** and uses a single classification model for multiple word recognition. We have tested this algorithm for recognition of all English digits on a speaker dependent model.

Algorithm Description

- **Viola Jones:** used for mouth/lip region detection and extraction.

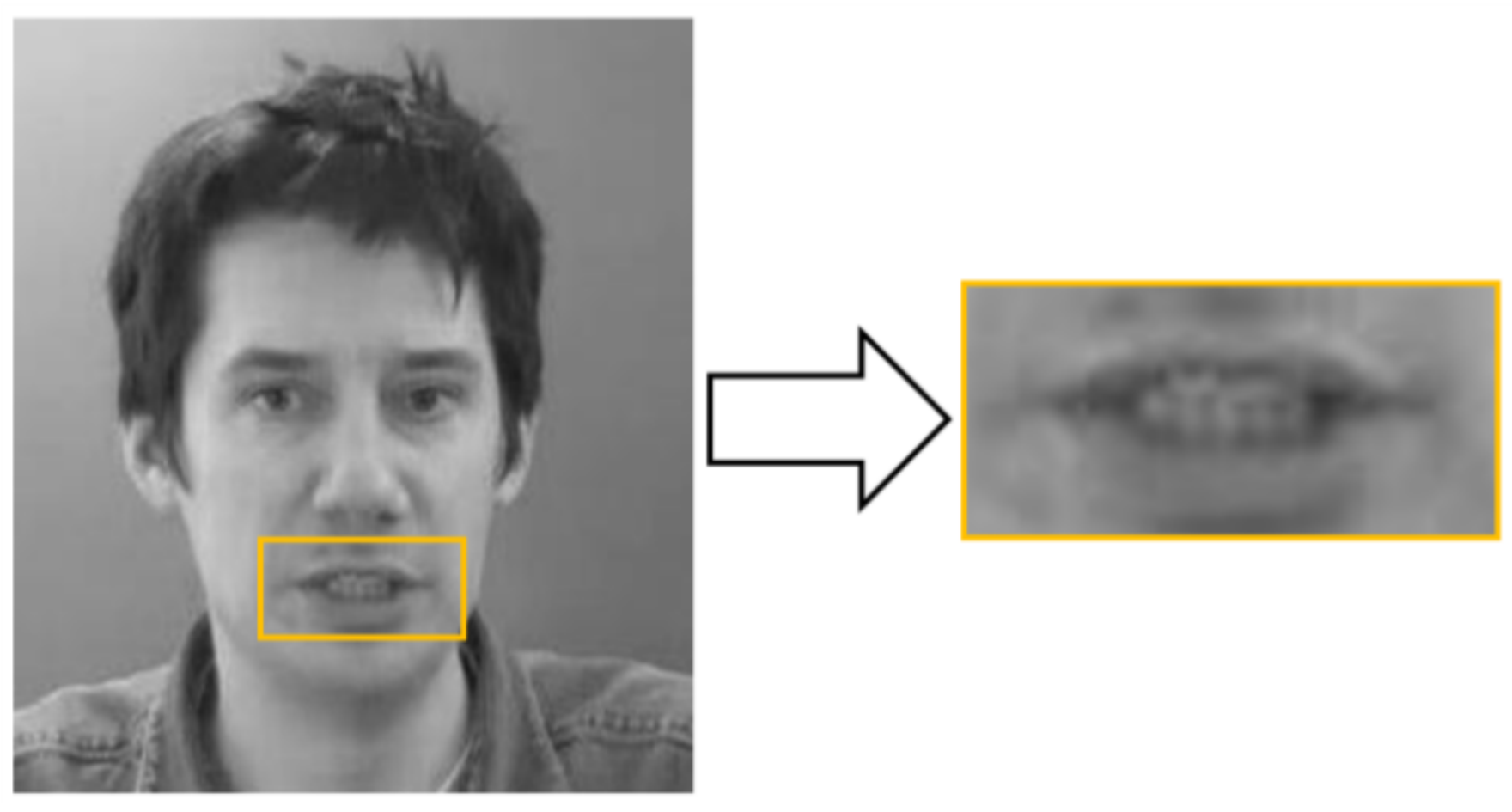


Figure 3: Lip region Extraction using Viola Jones algorithm

- **Local Binary Pattern:** used as a visual descriptor for each frame.

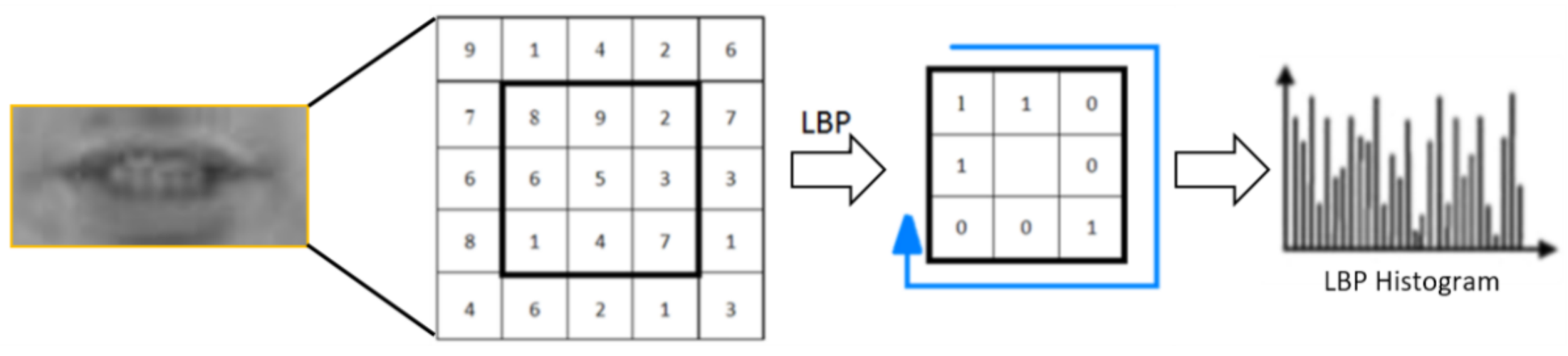


Figure 4: LBP Histogram from Lip Image

- **Discrete Cosine Transform:** used to convert the concatenated LBP histograms image into a 1D feature vector. Zig-zag scan of the DCT coefficients is done to get the 1D vector.
- **Support Vector Machine:** used for classification of feature vectors. A multi-class SVM model is trained for each subject. Each class corresponds to one digit.

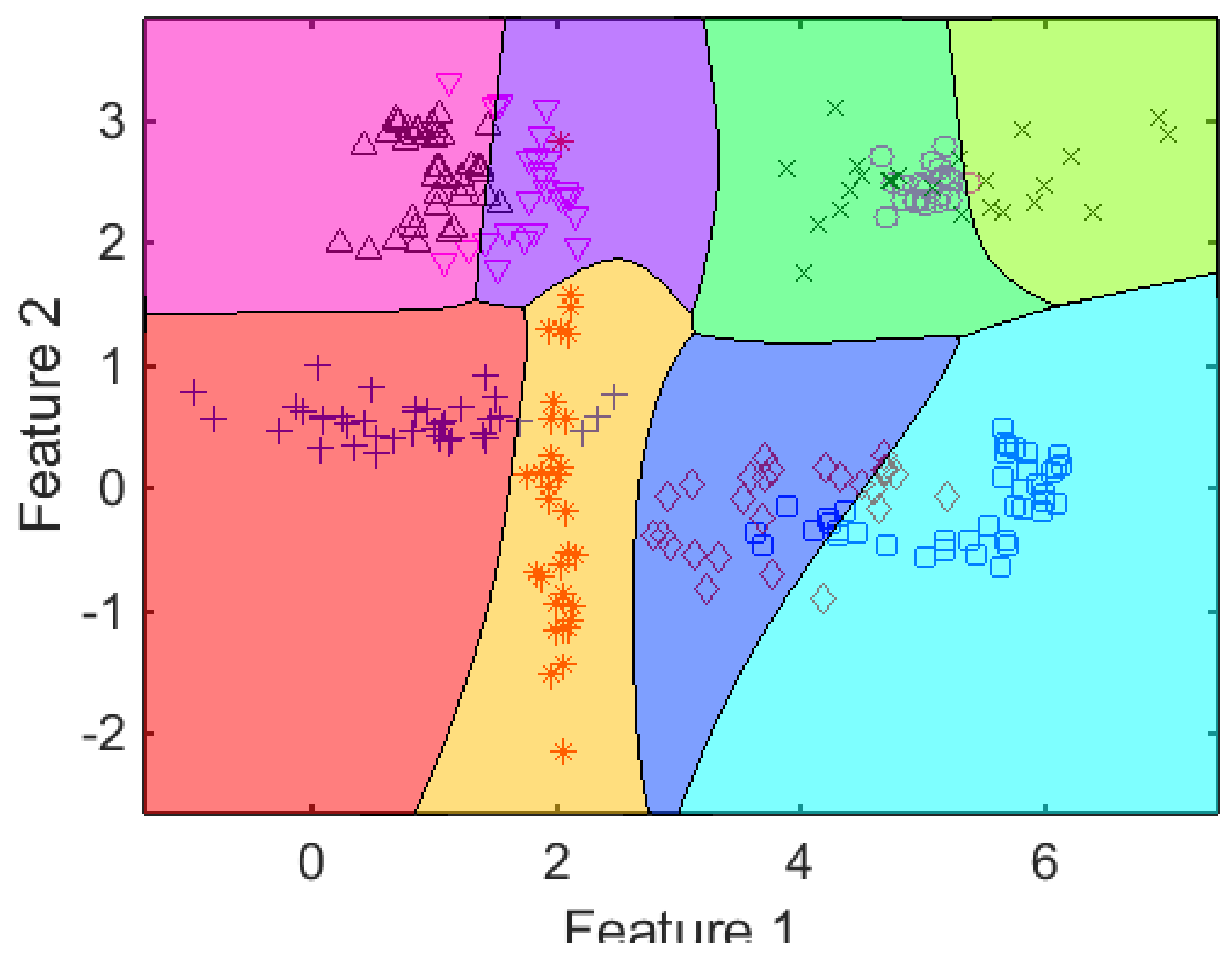
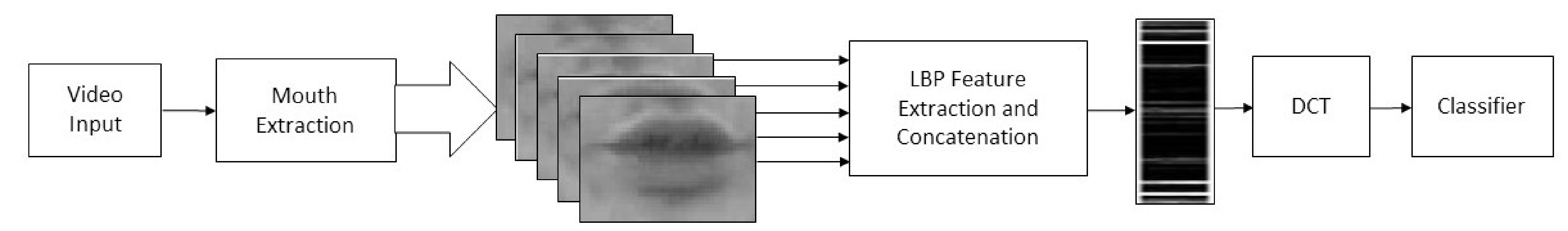


Figure 5: Multi-class SVM Model

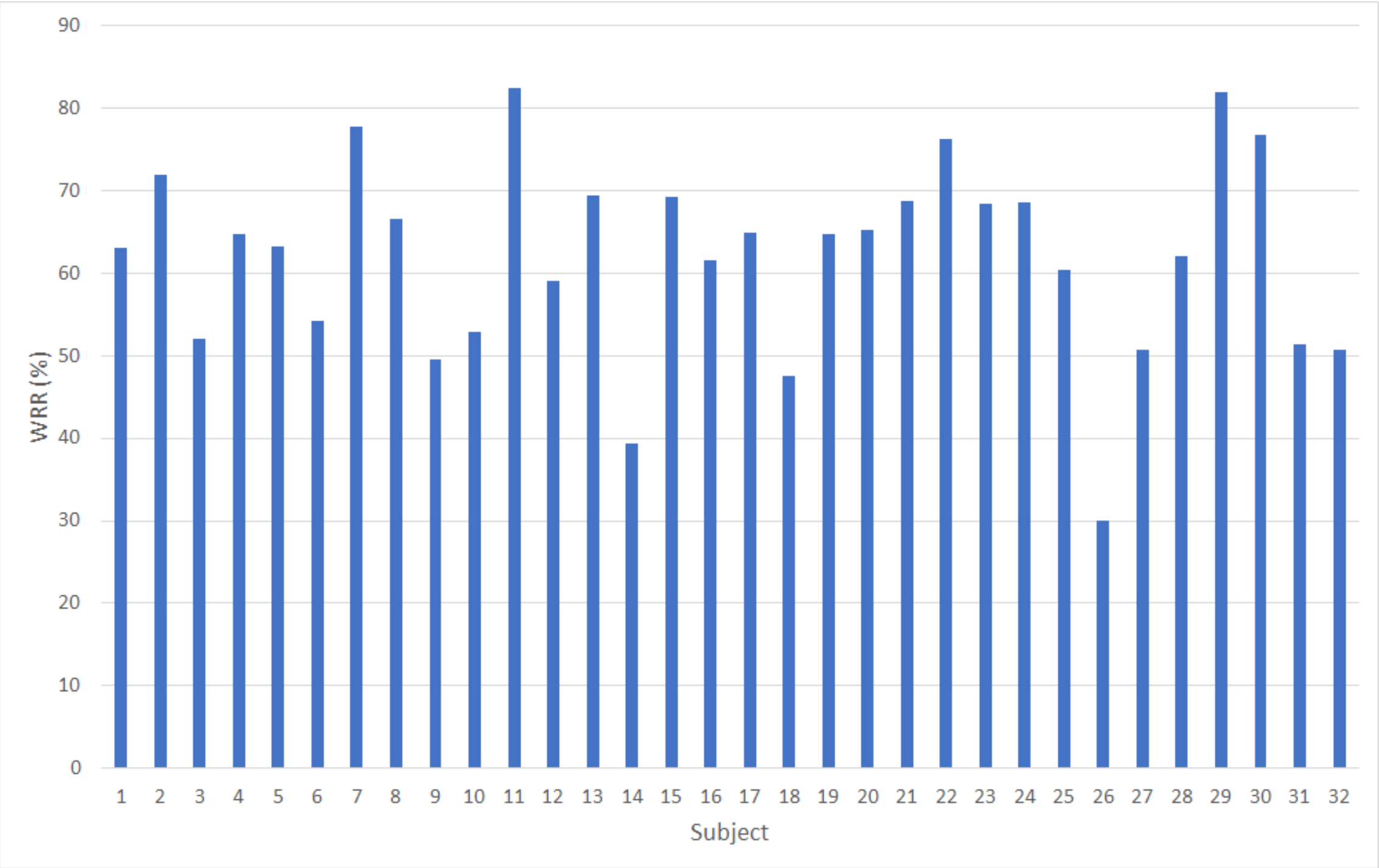
Algorithm Flowchart



Video Input → Mouth Extraction → LBP Feature Extraction and Concatenation → DCT → Classifier

Experiments and Results

Dataset used: GRID Audio-Visual sentence corpus [2]. The bar chart below shows the average Word Recognition Rate for each of the subjects. Overall average accuracy obtained is **62.02%**.



Contributions

- Computationally cheaper feature extraction
- Feature extraction requires no prior training
- Can be parallelized for each frame
- Single classification model for larger vocabulary

Conclusion

We presented a new algorithm for visual speech recognition. The proposed algorithm does not use any lip segmentation or modeling techniques, like ASM. The set of visual features is constructed using LBP, followed by DCT. For classification, SVM models are used. Average word recognition rate of 62.02% shows that there is definite scope for improvement. However, it demonstrates that relevant *visual features can be extracted without any lip modeling*, which is one of the most challenging tasks in such a system. Also, only a single SVM model was used for recognition of multiple words.

References

[1] S.L. Wang, A.W.C. Liew, W.H. Lau, and S.H. Leung. An Automatic Lipreading System for Spoken Digits with Limited Training Data. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 18, pages 1760–1765. IEEE, 2008.

[2] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *Acoustical Society of America*, pages 2421–2424, 2006.