ABSTRACT
Visual Speech Recognition (VSR) deals with the task of extracting speech information from visual cues from a person’s face while speaking. Accurate lip segmentation and modeling are essential in any VSR algorithm for good feature extraction. However, lip modeling is a complicated task and is not very robust in natural conditions. This paper describes a novel technique for limited vocabulary visual-only speech recognition that does not use lip modeling. For visual feature extraction, Discrete Cosine Transform (DCT) and Local Binary Pattern (LBP) have been tested. An Error-Correcting Output Codes (ECOC) multi-class model using Support Vector Machine (SVM) binary learners is used for recognition and classification of words.

Index Terms—visual speech recognition, local binary patterns, discrete cosine transform, feature extraction

1. INTRODUCTION
Lip reading is used to interpret and understand speech without hearing it. A technique mostly used by people with hearing disabilities. However, it has been shown that incongruent visual cues, paired with a particular auditory sound can cause people to perceive the sound or syllable in an entirely different manner. This phenomenon is called the McGurk effect [2], and it demonstrates the fact that visual cues contain relevant speech information.

Automatic lip reading or Visual Speech Recognition (VSR) is the task of extracting speech information just from visual inputs. The prime application of VSR systems is to provide complementary information to Audio Speech Recognition (ASR) systems for better accuracy. However, in scenarios where the audio input is highly corrupted by noise or is not present at all, speech detection solely depends on the visual inputs, like a video of a person talking.

Extracting relevant information from this visual inputs is one of the most important and difficult tasks in VSR. Other challenges faced in building a VSR system are robust face and mouth detection, accurate lip contour modeling (for feature extraction), and appropriate classifier design.

Geometric-, appearance-, and image-transform-based approaches are the different features extraction methods used currently [3]. Geometric- and appearance-based features use techniques like Active Shape Model (ASM) and Active Appearance Model (AAM) for lip segmentation and modeling [4]. Image-transform based features make use of transformation techniques like the Discrete Fourier (DFT), Discrete Wavelet (DWT) or Discrete Cosine (DCT) Transform for relevant feature extraction.

For classification, Hidden Markov Model (HMM), with Gaussian mixture observation densities, is widely used [4]. HMM implementation in VSR is similar to its implementation in ASR systems. Each HMM model identifies one word, with each state in the model describing a particular viseme. Other classification methods used in the literature are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Dynamic Time Warping (DTW) or some combination of these [5].

The proposed algorithm uses lip-image sequences for isolated digit (0-9) recognition. Viola-Jones algorithm [6] is used for mouth region extraction from the video frames. Two different types of features, namely DCT coefficients and LBP histogram, were tested. Both these features are computed over the region of interest (ROI) of each frame and concatenated to form a grayscale image. DCT was again used on the obtained grayscale image for final feature extraction. Multi-class Error-Correcting Output Codes (ECOC) [12] models were used for testing. Both speaker-dependent (SD) and speaker-independent (SI) models were tested.

2. ALGORITHM DESCRIPTION
This section describes the proposed algorithm in details. Since the algorithm tackles the visual-only speech recognition task, no audio signal was used throughout the process.

First, the video input of a person saying one of the English digits is taken. Lip region extraction is done on all video frames to obtain a lip-image sequence for that video. From each image in the sequence, one-dimensional (1D) feature vectors are extracted. The obtained vectors are concate-
nated to form a two-dimensional (2D) feature matrix, which can also be treated as a grayscale image. Two different feature extraction methods – DCT and LBP – were used to create the feature matrix. Both types of features were tested independently.

After the grayscale image is created, DCT is again used to convert it into the final 1D feature vector, which is used to train the ECOC classifier. Fig. 1 shows the entire process described above.

2.1. Mouth/Lip Region Detection
Viola-Jones [6] is a robust and efficient object detection framework. It is used here for mouth/lip region extraction. Fig. 2 shows a frame from the dataset with the mouth detected using Viola-Jones algorithm. The yellow bounding box shows the region of interest (ROI) for feature extraction. Since the mouth is the part of the face which contains the most relevant speech information, feature extraction is done on just that part. ROI is cropped from all the video frames, and further processing is done on the cropped frames.

Feature extraction in the proposed algorithm falls under the category of image transform based approach. After ROI extraction using Viola-Jones in the previous step, LBP and DCT features are computed. Both these algorithms do not need any prior training which makes a simpler feature extraction method. Another advantage of using LBP/DCT is that it is highly parallelizable and recognition can be faster.

2.2. Feature Extraction
Most of the existing VSR algorithms use some lip-modeling technique, like ASM or AAM, for lip contour detection or locating corner points of the mouth, for geometric features extraction. This task needs some additional training to correctly create a lip model and identify it on the test images. Moreover, in natural conditions, the results may be inaccurate [7].

2.2.1. LBP/DCT
LBP is generally used as a feature for texture detection [8]. It has also been used as a feature descriptor in algorithms like face detection [9]. In the proposed algorithm, LBP histogram is used as a feature vector. The described process is shown in Fig. 3.

![Fig. 3. LBP histogram](image)

DCT has been extensively used as a data compression technique since it’s introduction in 1974 [10]. DCT has also been used as a feature for VSR in the literature [7] & [11]. However, it has mostly been used in conjunction with other dimensionality reduction/feature extraction methods like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). Lip modeling is still used in these algorithms for ROI detection.

Motivated by these, the proposed algorithm also employs DCT as a feature, in the same way as LBP described above. Zig-zag scan of the DCT coefficients is done in order to obtain a 1D feature for each image. Only the first 200 DCT coefficients are used.

2.2.2. Final Feature Vector
Once the feature matrix is obtained, DCT is again applied to get the final feature vector. Zig-zag is done and only the first 200 DCT coefficients after a zig-zag scan are used. This process is shown in Fig. 4.
2.3. Classification

The most popular classification technique for ASR and VSR systems is HMM [4]. However, in the holistic approach of VSR (recognizing entire words instead of each syllable), one HMM needs to be trained for each word. In order to avoid using multiple models, a single model for multi-class classification is used.

ECOC is a method designed to tackle multi-class classification tasks [12]. It combines multiple binary classifiers to solve a multi-class problem.

SVM is a supervised machine learning technique which uses separating hyperplanes to distinguish between two classes of data, essentially making it a binary classifier. Hence, it is used as a binary learner for ECOC model implementation.

There are many coding schemes to design an ECOC model, like one-versus-one, one-versus-all, ordinal, etc. For the proposed algorithm, one-versus-one coding design is used, which creates $\frac{k(k-1)}{2}$ SVMs per ECOC model, where $k$ is the number of classes. For the digit classification task, $k = 10$, which gives 45 binary learners per model.

3. EXPERIMENTAL EVALUATION

3.1. GRID Audio-Visual Sentence Corpus

The dataset used for evaluation of the algorithm is GRID audio-visual sentence corpus [13]. It contains 1000 videos each, of 34 speakers (16 male and 18 female). Each video is of the person speaking a simple 6-word sentence, containing a one-digit number, for example, “place green at B 4 now.” The videos were recorded at 25 frames per second, and have been made available in two resolutions: normal (360x280) and high (720x576). Normal resolution videos were used for experimentation. A sample frame from the dataset (converted to grayscale) can be seen in Fig. 2.

Along with each video file, an alignment file is also made available. This file contains the beginning and ending time-frames of each of the six words spoken in the video. Using this information, the video frames where digits were spoken, were extracted and used.

3.2. Results

Isolated digit recognition for all digits (0-9) was tested using two different feature extraction methods. Out of 34 [13], the dataset for 30 speakers was used. This gives a total of 30,000 videos; with 3,000 videos per digit for evaluation of the algorithm. Both speaker-dependent and independent models were tested.

For the speaker dependent (SD) case, each model was trained on randomly selected 90% of the feature set, with the remaining 10% being used for testing. A hundred iterations of this testing process were done, and the average Word Recognition Rate (WRR) for each speaker is shown in Fig. 5. The average WRR, for speaker dependent models, over all speakers is 70.68% for LBP features, and 79.72% for DCT features. It is apparent from the results that DCT features give better accuracy compared to LBP features in most cases.
For speaker independent (SI) testing, five models were trained. From the dataset of 30 speakers, 25 were randomly selected for creating the models, and the other 5 were used as the test set. The 25 speakers used for training were randomly divided into five groups of five speakers each. One model per group was trained.

Having multiple models enables capturing a wider range of speaking styles; this means that for a particular speaker, one model can perform much better than all others, which is reflected in the results shown in Fig. 6 & 7.

The average accuracy for speaker independent models is calculated using just the maximum WRR obtained for a speaker across all models. This gives an average of 48.8% for DCT features, and 33.5% for LBP.

These results are comparable to a recently introduced algorithm for isolated digit VSR [14], which uses lip modeling along with DCT for feature extraction, and HMM as a classifier. The video-only deep auto-encoder technique introduced in [15] focuses on uses deep learning for audio-visual speech recognition. The results used here are for visual-only speech recognition for isolated digits. The authors in [16] have again used lip-modeling and one HMM model per word for classification. A summary of these results is given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>SD WRR (%)</th>
<th>SI WRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT</td>
<td>79.72</td>
<td>48.8</td>
</tr>
<tr>
<td>(Proposed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBP</td>
<td>70.68</td>
<td>33.5</td>
</tr>
<tr>
<td>(Proposed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCT+Lip Model[14]</td>
<td>80.78</td>
<td>52.93</td>
</tr>
<tr>
<td>Video-only Deep</td>
<td>-</td>
<td>68.7</td>
</tr>
<tr>
<td>Auto-Encoder[15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zernike+ PCA[16]</td>
<td>-</td>
<td>63.88</td>
</tr>
</tbody>
</table>
5. REFERENCES


