# VISUAL SPEECH RECOGNITION USING WAVELET TRANSFORM AND MOMENT BASED FEATURES

Wai C. Yau, Dinesh K. Kumar, Sridhar P. Arjunan and Sanjay Kumar
*School of Electrical and Computer Engineering*
*RMIT University, GPO Box 2476V Melbourne, Victoria 3001, Australia*

Keywords: Visual Speech Recognition, Motion History Image, Discrete Stationary Wavelet Transform, Image Moments, Artificial Neural Network.

Abstract: This paper presents a novel vision based approach to identify utterances consisting of consonants. A view based method is adopted to represent the 3-D image sequence of the mouth movement in a 2-D space using grayscale images named as motion history image (MHI). MHI is produced by applying accumulative image differencing technique on the sequence of images to implicitly capture the temporal information of the mouth movement. The proposed technique combines Discrete Stationary Wavelet Transform (SWT) and image moments to classify the MHI. A 2-D SWT at level 1 is applied to decompose MHI to produce one approximate and three detail sub images. The paper reports on the testing of the classification accuracy of three different moment-based features, namely Zernike moments, geometric moments and Hu moments computed from the approximate representation of MHI. Supervised feed forward multilayer perceptron (MLP) type artificial neural network (ANN) with back propagation learning algorithm is used to classify the moment-based features. The performance and image representation ability of the three moments features are compared in this paper. The preliminary results show that all these moments can achieve high recognition rate in classification of 3 consonants.

## 1 INTRODUCTION

### Lip Reading for Human Computer Interface

With the rapid advancement in Human Computer Interaction (HCI), speech recognition has become one of the key areas of research among the computer science and signal processing community. Speech driven interfaces are being developed to replace the conventional interfaces such as keyboard and mouse to enable users to communicate with the computers using natural speech. However, these systems are based on audio signals and are sensitive to signal strength, ambient noise and acoustic conditions.

The human speech production and perception system is known to be bimodal and consists of the audio modality and the visual modality(Chen, 2001). The visual speech information refers to the movement of the speech articulators such as the tongue, teeth and lips of the speaker. The complex range of reproducible sounds produced by people is a clear demonstration of the dexterity of the human mouth and lips-the key speech articulators. This project proposes the use of video data related to lip and mouth movement for human computer interface applications. The possible advantages are that such a system is not sensitive to audio noise and change in acoustic conditions, does not require the user to make a sound, and provides the user with a natural feel of speech and the dexterity of the mouth. Such a system is termed by the authors as 'Audio-less Speech Recognition' system.

Audio-less speech recognition requires using only the sensing of the facial movement. There are a number of options that have been proposed, such as visual, mechanical sensing of facial movement and movement of palate, recording facial muscle activity(Kumar et al., 2004) and facial plethysmogram. Speech recognition based on visual speech signal is the least intrusive and this paper reports such a system for human computer interface applications.The visual cues contain far less classification power for speech compared to audio data (Potamianos et al., 2003) and hence it is to be expected that the visual only systems would have only a small vocabulary.

Visual speech recognition techniques reported in the literature in the past decade can be catego-



C. Yau W., K. Kumar D., P. Arjunan S. and Kumar S. (2006).
VISUAL SPEECH RECOGNITION USING WAVELET TRANSFORM AND MOMENT BASED FEATURES.
In *Proceedings of the Third International Conference on Informatics in Control, Automation and Robotics*, pages 340-345
DOI: 10.5220/0001209903400345

rized into two main categories -shape-based and the appearance-based. The shape-based features rely on the geometric shape of the mouth and lips and can be represented by a small number of parameters. One of the early visual speech recognition system was developed by Petajan(Petajan, 1984) using shape-based features such as height, width and area of the mouth from the binary image of the mouth. Appearance-based features are derived directly from the pixel intensity values of the image around the mouth area (Liang et al., 2002; Potamianos et al., 2003).

One common difficulty with visual systems is that these systems are 'one size fits all' approach. Due to the large variation in the way people speak, especially if we transgress the national and cultural boundaries, these have very high error rate, with error of the order of 90% (Potamianos et al., 2003). This demonstrates the inability of these systems to be used for such applications. What is required is a system that is easy to train for a user, which works in real-time and be robust under changing conditions (such as position of the camera, speed of the speech and skin color.)

To achieve the above mentioned goals, this paper proposes a system where the camera is attached in place of the microphone to the commonly available head-sets. The advantage of this is that using this, it is no longer required to identify the region of interest, reducing the computation required. The video processing proposed is the use of accumulative image differencing technique based on the use of motion history image (MHI) to directly segment the movement of the speech articulators. MHI is invariant to factors such as the skin color and texture of the speakers. This paper reports the use of 2-D stationary wavelet transform (SWT) at level 1 to decompose the MHI into four sub images, with the approximate image used for further analysis. This paper proposes to use image moments as features extracted from the approximation of MHI and ANN to classify these features. The fundamental concept of this technique can be traced to the research reported by Bobick and Davis(Bobick and Davis, 2001) in classifying human movement and the work by Kumar et. al(Kumar and Kumar, 2005) in hand gesture recognition and biometrics identification.

# 2 THEORY

## 2.1 Motion History Image

Motion history image (MHI) is a view-based approach and is generated using difference of frames (DOF) from the video of the speaker. Accumulative image difference is applied on the image sequence by subtracting the intensity values between successive frames to generate the difference of frames (DOFs). The delimiters for the start and stop of the motion are manually inserted into the image sequence of every articulation. The MHI of the video of the lips would have pixels corresponding to the more recent mouth movement brighter with larger intensity values.

The intensity value of the MHI at pixel location (x, y) of the $t^{th}$ frame is defined by

$$MHI_t = max \bigcup_{t=1}^{N-1} B(x,y,t) \times t \qquad (1)$$

N is the total number of frames used to capture the mouth motion. n represents the B(x,y,t) is the binarisation of the DOF using the threshold a and B is given by

$$B(x,y,t) = \begin{cases} 1 & \text{if } Diff(x,y,t) \geq a, \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

a is the predetermined threshold for binarisation of the DOF and

$$Diff(x,y,t) = |I(x,y,t) - I(x,y,t-1)| \qquad (3)$$

I(x, y, t) represents the intensity value of pixel location with coordinate (x, y) at the $t^{th}$ frame of the image sequence. Diff(x, y, t) is the DOF of the $t^{th}$ frame. In Eq. (1), the binarised version of the DOF is multiplied with a linear ramp of time to implicitly encode the timing information of the motion into the MHI. By computing the MHI values for all the pixels coordinates (x, y) of the image sequence using Eq. (1) will produce a scalar-valued grayscale image (MHI) where the brightness of the pixels indicates the recency of motion in the image sequence(Kumar and Kumar, 2005).

The motivation of using MHI in visual speech recognition is the ability of MHI to remove static elements from the sequence of images and preserve the short duration complex mouth movement. MHI is also invariant to the skin color of the speakers due to the DOF process. Further, the proposed motion segmentation approach is computationally simple and is suitable for real time implementation.

MHI is a view sensitive motion representation technique. Therefore the MHI generated from the sequence of images of different consonants is dependent on factors such as:

- position of the speaker's mouth normal to the camera optical axis
- orientation of the speaker's face with respect to the video camera
- distance of the speaker's mouth from the camera
- small variation of the mouth movement of the speaker while uttering the same consonant

It is difficult to ensure that the position, orientation and distance of the speaker's face are constant from the video camera for every sample taken. Thus, descriptors that are invariant to translation, rotation and scale have to be used to represent the MHI for accurate recognition of the consonants. The features used to describe the MHI should also be insensitive to small variation of mouth movement between different samples of the same consonants. This paper adopts image moments as region-based features to represent the approximation of the MHI. Image moments are chosen because they can be normalized to achieve scale, translation and rotation invariance. Before extracting the moment-based features, SWT is applied to MHI to obtain a transform representation of the MHI that is insensitive to small variations of the mouth and lip movement.

## 2.2 Stationary Wavelet Transform (SWT)

2-D SWT is used for denoising and to minimize the variations between the different MHI of the same consonant. While the classical discrete wavelet transform (DWT) is suitable for this, DWT results in translation variance (Mallat, 1998) where a small shift of the image in the space domain will yield very different wavelet coefficients. SWT restores the translation invariance of the signal by omitting the downsampling process of DWT, and results in redundancies.

2-D SWT at level 1 is applied on the MHI to produce a spatial-frequency representation of the MHI. SWT decomposition of the MHI generates four images, namely approximation (LL), horizontal detail coefficients (LH), vertical detail coefficients (HL) and diagonal detail coefficients (HH) through iterative filtering using low pass filters H and high pass filters G. The approximate image is the smoothed version of the MHI and carries the highest amount of information content among the four images. LH, HL and HH sub images show the fluctuations of the pixel intensity values in the horizontal, vertical and diagonal directions respectively. The image moments features are computed from the approximate sub image.

## 2.3 Moment-based Features

Image moments are low dimensional descriptors of image properties. Image moments features can be normalized to achieve translation, rotation and scale invariance(Mukundan and Ramakrishnan, 1998) thus are suitable to be used as features to represent the approximation of MHI.

**Geometric moments**

Geometric moments are the projection of the image function f(x, y) onto a set monomial function.The regular geometric moments are not invariant to rotation, translation and scaling.

Translation invariance of the features can be achieved by placing the centroid of the image at the origin of the coordinate system (x, y), this results in the central moments. The central moments can be further normalized to achieve scale invariant. The normalized central moments are invariant to changes in position and scale of the mouth within the MHI.

The normalized central moments can be derived up to any order. In this paper, the 49 normalized geometric moments up through 9th order are computed from the MHI as one of the feature descriptors to represent the different consonants. For the purpose of comparison of the different techniques, the total number of moments has been kept the same. Zernike moments require 49 moments, and thus this number has been kept for the geometric moments as well.

**Hu moments**

Hu (Hu, 1962) introduced seven nonlinear combinations of normalized central moments that are invariant to translational, scale and rotational differences of the input patterns known as absolute moments invariants. The first six absolute moment invariants are used in this approach as features to represent the approximate image of the MHI for each consonant. The seventh moment invariant is skew invariant defined to differentiate mirror images and is not used because it is not required in this application.

**Zernike moments**

Zernike moments are computed by projecting the image function f(x, y) onto the orthogonal Zernike polynomial. The main advantage of Zernike moments is the simple rotational property of the features. Zernike moments are also independent features due to the orthogonality of the Zernike polynomial(Teague, 1980). This paper uses the absolute value of the Zernike moments as the rotation invariant features(Khontazad and Hong, 1990) of the SWT of MHI. 49 Zernike moments that comprise of 0th order moments up to 12th order moments have been used as features to represent the approximate image of the MHI for each consonant.

## 2.4 Classification using Artificial Neural Network

Classification involves assigning of new inputs to one of a number of predefined discrete classes.

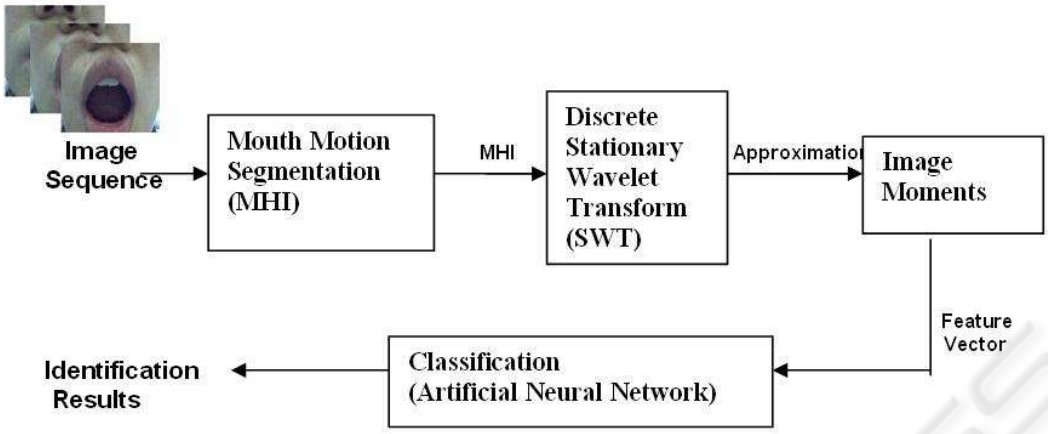


Figure 1: Block Diagram of the proposed visual speech recognition approach.

There are various classifier choices for pattern recognition applications such as Artificial Neural Network (ANN), Bayesian classifier and Hidden Markov Model (HMM). In this paper, we present the use of ANN to classify moment-based feature input into one of the class of viseme. ANN has been selected because it can solve complicated problems where the description for the data is not easy to compute. The other advantage of the use of ANN is its fault tolerance and high computation rate due to the massive parallelism of its structure (Kulkarni, 1994). The functionality of the ANN to be less dependent on the underlying distribution of the classes as opposed to other classifiers such as Bayesian classifier is yet another advantage for using ANN in this application.

A supervised feed-forward multilayer perceptron (MLP) ANN classifier with back propagation learning algorithm is integrated in the visual speech recognition system described in this paper. The ANN is provided with number of training vectors for each class during the training phase. MLP ANN was selected due to its ability to work with complex data compared with a single layer network. Due to the multilayer construction, such a network can be used to approximate any continuous functional mapping (Bishop, 1995). The advantage of using back propagation learning algorithm is that the inputs are augmented with hidden context units to give feedback to the hidden layer and extract features of the data from the training events (Haung, 2001). Trained ANNs have very fast classification speed thus making them an appropriate classifier choice for real time visual speech recognition applications.Figure 1 shows the block diagram of the proposed system.

# 3 METHODOLOGY

Experiments were conducted to evaluate the performance of the proposed visual speech recognition approach. The experiments were approved by the Human Experiments Ethics Committee of the University. The experiment was designed to test the efficiency of different moments features in classifying 3 consonants when there was none or minimal shift between the camera and the mouth of the user between the training and the testing data.

The first step of the experiment was to record the video data from a speaker uttering the three consonants. The three consonants selected were a fricative /v/, a nasal /m/ and a stop /g/. Each consonant was repeated for 20 times while the mouth movement of the speaker was recorded using an inexpensive web camera. This was done towards having an inexpensive speechless communication system using low resolution video recordings. The video camera focused on the mouth region of the speaker and the camera was kept stationary throughout the experiment with a constant window size and view angle. A consistent background and illumination was maintained in the experiments. The video data was stored as true color (.AVI) files and every AVI file had a duration of 2 seconds to ensure that the speaker had sufficient time to utter each consonant. The frame rate of the AVI files was 30 frames per second which is within the range of standard frame rate for video cameras. One MHI was generated from each AVI file. A total of 60 MHI were produced for the 3 consonants, 20 for each consonant. Example of the MHI for each of the consonants /v/, /m/ and /g/ are shown in Figure 2.

SWT at level-1 using Haar wavelet was applied on the MHI and the approximate image was used for

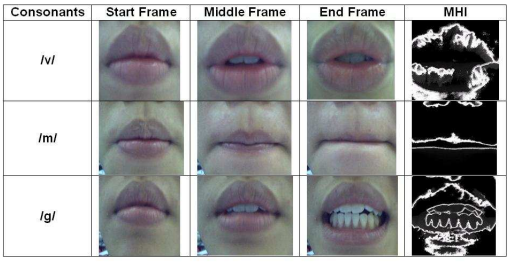


Figure 2: The three consonants and the MHI for the consonants.

analysis. The moment-based features were extracted to characterize the different mouth movement of the approximate image of MHI generated by the SWT. 49 moments -each of Zernike, geometric and first six Hu moments were computed. These features were tested to determine the efficiency of the different moments in representing the lip movement.

In the experiment, features from 10 MHI (for each consonant) were used to train the ANN classifier with back propagation learning algorithm. The architecture of the ANN consisted of two hidden layers and the numbers of nodes for the two hidden layers were optimized iteratively during the training of the ANN. Sigmoid function was the threshold function and the type of training algorithm for the ANN was gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima. The trained ANNs were tested by classifying the remaining 10 MHI of each consonant that were not used in the training of the ANN to test the performance of the proposed approach. The performance of these three moment-based features was evaluated in this experiment by comparing the accuracy in the classification during testing.

## 4 RESULTS AND OBSERVATIONS

The experiment investigates the performance of the different features in classifying the MHI of the 3 consonants. The classification results are tabulated in Table 1. It is observed that the classification accuracies of the three features (Zernike moments, geometric moments and Hu moments) are very similar, with Zernike moments and geometric moments yielding marginally higher recognition rate (100%) compared to Hu moment features. The results also indicate a very high level of accuracy for the system to identify the spoken consonant from the video data when there is no relative shift between the camera and the mouth.

Table 2 summarizes the recognition rates for the three moment features. The results indicate that the MHI based technique is able to recognize spoken consonants with a high degree of accuracy.

## 5 DISCUSSION

The results indicate that the technique provides excellent results for identifying the unspoken sound based on video data, with error less than 5%. The results using the three different image moments are all very comparable. Hu moment has marginally higher error, but has significantly smaller number of moments used (only 6 moments) to represent the MHI compared with Zernike moments and geometric moments, which have 49. The higher order moments contain more information content of the MHI (Teh and Chin, 1988).

While this error rate is far lower than the 90% error reported in the review by (Potamianos et al., 2003), the authors believe that it is not appropriate to compare the work reported there to our work as this system has only been tested for limited and selected phones.

The promising results obtained in the experiment indicate that this approach is suitable for classifying consonants based on the mouth movement without regard to the static shape of the mouth. The results also demonstrate that a computationally inexpensive system which can easily be developed on a DSP chip can be used for such an application. At this stage, it should be pointed that this system is not being designed to provide the flexibility of regular conversation language, but for a limited dictionary only, and where the phones are not flowing, but are discrete. The current systems require the identification of the start and end of the phone, and the authors propose the use of muscle activity sensors for this aim.

The authors would also like to point out that this system is part of the overall system. This system is designed for consonant type of sounds, where there is facial movement during the phone. The authors have also designed a separate system that is suitable for vowels, and the two need to be merged together for the complete system.

The authors believe that one reason for better results of this system compared with the other related works is that it is not only based on lip movement, but is based on the movement of the mouth. While lips are important articulators of speech, other parts of the mouth are also important, and this approach is closer to the accepted speech models.

Table 1: Classification results for different image moments extracted from the SWT approximate image of the MHI.

| Actual Consonants | Predicted Consonants | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Zernike | | | Geometric | | | Hu | | |
| | /v/ | /m/ | /g/ | /v/ | /m/ | /g/ | /v/ | /m/ | /g/ |
| /v/ | 10 | - | - | 10 | - | - | 10 | - | - |
| /m/ | - | 10 | - | - | 10 | - | 1 | 9 | - |
| /g/ | - | - | 10 | - | - | 10 | - | - | 10 |

Table 2: Recognition rates for the different moment features of the MHI.

| Type of Moments | Recognition Rate |
|---|---|
| Zernike Moments | 100% |
| Geometric Moments | 100% |
| Hu Moments | 96.67% |

# 6 CONCLUSION

This paper describes a visual speech recognition approach that is based on direct mouth motion representation and is suitable for real time implementation. The low complexity of the proposed visual speech recognition system is achieved by using image differencing technique that represent the mouth motion of the image sequence using grayscale images, motion history image (MHI).

This paper focused on classifying English consonants because pronunciation of consonants results in more visually observable movement of the speech articulators as compare to vowels. 2-D SWT and image moments (Zernike moments, geometric moments and Hu moments) are used to extract visual speech features from the MHI and classification of these features is performed by ANN. The experimental results indicate that such a system can produce high classification rate (approximately 100%) using moment based features extracted from SWT approximate of MHI.

There is a need to identify the limitation of this technique by testing the system on large number of sounds, with the intent to determine the possible vocabulary that maybe supported by such a technique. Such a system could be used to drive computerized machinery when in noisy environments. The system may also be used for helping disabled people to use a computer and for voice-less communication.

# REFERENCES


Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267.

Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18:9–21.

Haung, K. Y. (2001). Neural network for robust recognition of seismic patterns. In *IIJCNN'01, Int Joint Conference on Neural Networks*.

Hu, M. K. (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8:179–187.

Khontazad, A. and Hong, Y. H. (1990). Rotation invariant image recognition using features selected via a systematic method. *Pattern Recognition*, 23:1089–1101.

Kulkarni, A. D. (1994). *Artificial Neural Network for Image Understanding*. Van Nostrand Reinhold.

Kumar, S. and Kumar, D. K. (2005). Visual hand gesture classification using wavelet transform and moment based features. *International Journal of Wavelets, Multiresolution and Information Processing(IJWMIP)*, 3(1):79–102.

Kumar, S., Kumar, D. K., Alemu, M., and Burry, M. (2004). Emg based voice recognition. In *Intelligent Sensors, Sensor Networks and Information Processing Conference*.

Liang, L., Liu, X., Zhao, Y., Pi, X., and Nefian, A. V. (2002). Speaker independent audio-visual continuous speech recognition. In *IEEE Int. Conf. on Multimedia and Expo*.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press.

Mukundan, R. and Ramakrishnan, K. R. (1998). *Moment Functions in Image Analysis : Theory and Applications*. World Scientific.

Petajan, E. D. (1984). Automatic lip-reading to enhance speech recognition. In *GLOBECOM'84, IEEE Global Telecommunication Conference*.

Potamianos, G., Neti, C., Gravier, G., and Senior, A. W. (2003). Recent advances in automatic recognition of audio-visual speech. In *Proc. of IEEE*, volume 91.

Teague, M. R. (1980). Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70:920–930.

Teh, C. H. and Chin, R. T. (1988). On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:496–513.