International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 26 Issue 5 – AUGUST 2019.

Lip Event Recognition and Geometric Feature Extraction for Lip Reading System

Thein Thein [#1] and Kalyar Myo San [*2]

[#] *Faculty of Information Science, University of Computer Studies (Mandalay), Myanmar*

[*] *Faculty of Computer Systems and Technologies, University of Computer Studies (Mandalay), Myanmar*

***Abstract*—A lip reading system is a communication technique used by a hearing person in a conversation. Now and again, the word they understand does not match what the other speaker says. A lip reading system can make them trace these words based on the movements of the lips. Many algorithms and methods are proposed to recognize lip movement and to extract features from the movement of the lip. To recognize the spoken word, lip event detection and feature extraction is need. In this paper, lip event is detected by using a La*b* color space method and Moore Neighborhood Algorithm. Then, features are suggested as visual features of the motion of the lip based on geometrical information. The research goal in this study is to recognize lip motion based on modifications in the ellipse surface area. For the experiments, several spoken consonants have been chosen. The accuracy of proposed method is verified by using it to recognize 14 two syllable consonants of Myanmar Language.**

***Index Terms*— La*b* color space, Lip reading, Lip movements, Moore Neighborhood Algorithm.**

I. Introduction

Lip reading system is systems that can assist deaf and hard hearing people learn to talk with the correct motion of the lip. Due to the presence of noise in different circumstances, the additional use of visual characteristics is anticipated to enhance lip reading system efficiency. For both the method of motion recognition and visual feature extraction, automatic lip reading is difficult task. Extraction of the visual function needs a robust technique of monitoring the lips of the speaker through a series of images and a representation of the mouth. Lip monitoring is not a trivial task as there is a range of skin color, lip color, environmental variability such as lighting circumstances in individuals. The motion of the lips from frame to frame should also be adapted to any technique used to monitor the lips during speech. With regard to the method of recognition, various techniques have been created to identify the motion of the lip according to the visual characteristics.

Due to its attractive applications including lip reading [9], audio-visual speech recognition [10], facial expression analysis [11]-[12] and so on, lip event analysis in video has been widely researched in recent years. One of the main problems among these apps is the accurate detection of lip movement occurrences, so it is possible to obtain the respective lip dynamic state for lip behavior inquiry. Detecting lip dynamic states about the opening and closing of the mouth is essential to the assessment of the facial appearance.

This paper present a method for movement detection and lip features. Literature review is discussed in section 2. Methods for lip extraction are described in section 3. Moore Neighborhood Algorithm is introduced in section 4. In section 5, extracting lip features are presented to classify consonants. The proposed method is evaluated by employing it in recognition of 14 two syllable Myanmar consonants presented in section 6. Finally, we draw the conclusion and future works.

II. Literature Review

The gradient-based techniques were used by Delmas et al.[2] and Eveno et al.[3] to remove the lip border while the input image is regarded as a vector map. The precision of these techniques, however, is easily affected by fake border edges induced by shadow, skin pigmentation, etc. A. A linear discriminant analysis (LDA) is used by Nefian et al. [4] to distinguish the lip pixels from the skin pixels and thus to remove the contour of the lip. Although a smoothing procedure follows the LDA, the resulting segmentation is often loud.

The "snakes"[6] have been commonly applied to lip segmentation [5]-[7]-[8] due to their capacity to take into consideration smoothing and elasticity limitations. They can offer quite nice outcomes, but the tuning of parameters is very hard to accomplish most of the moment, and the snakes often converge when the original position is far from the edges of the lip. Tian [21] utilizes a straightforward parabola-based three-state geometric model. The data on color and shape is used to understand which model to use: closely closed, closed, or opened mouth. Then, the model is drawn using 4 main points. The model's position is generally good, but it doesn't match the limit with precision because it can only generate symmetrical parabolic forms. Other authors suggest using two parables instead of one for the upper boundary [22] or using quadratics instead of parabola [23] to create the model more flexible. It increases precision but, especially in the case of asymmetric teeth, the models are still restricted by their rigidity.

III. Lip Extraction

A. Lip Region Detection

To recognize lip movement and to extract lip feature, accurate lip region should be extracted. Detection of the region of the lip needed several image preprocessing processes.

The lip images are in RGB color space before subtracting the region of the lip. Fig.1 displays of the original lip images.

Selecting a color space is essential as it directly affects the robustness and precision of the segmentation. It is also possible to use CIELUV and YCrCb spaces for face analysis. The skin color subspace has been shown to cover a small area of planes (Cr, Cb) or (u, v)([24][25]). Skin and lip color distributions, however, often overlap and differ for distinct speakers. It makes these spaces unfit for the segmentation of the lips.

The initial color image is improved in the first phase by using decorrelation stretching color enhancement technique with stretch limit 0.5 to guarantee solid lip detection. The color improved image is shown in Fig.2. The RGB color image is then converted into Lab color space's first layer L channel. This is because the impact of lighting glare is avoided. Fig.3 demonstrates images that have been converted in color.

Fig.1 Original Images.

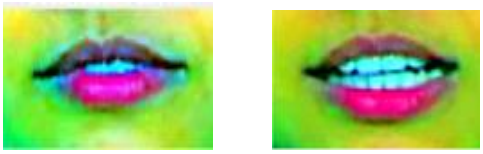

Fig.2 Enhanced Images.

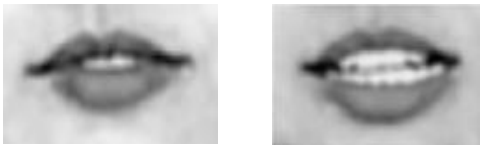

Fig. 3 Color Transformed Images.

B. Binary Transformation and Mouth Region Extraction

Methods aiming at segmenting the lip shape, boundary or mouth area from the images of the input video can be divided in two main types: region-based and model-based approaches. Region-based approaches try to find the mouth area only and use a rectangular box or ellipse around the mouth as segmentation output. Model-based approaches try to fit a certain shape model of the mouth to the data which results in finding the outer and/or inner boundary of the lips.

Simple image thresholding was used by Petajan[13] to obtain binary mouth images, height, perimeter, region and width as visual characteristics to create their speech reading system. Using threshold value to group its pixels into black and white is the easiest way to segment the lip image. For our studies, the transformed images are converted into binary images to obtain lip contour on the lip border. Fig. 4 shows the binary transformation on the lip image.

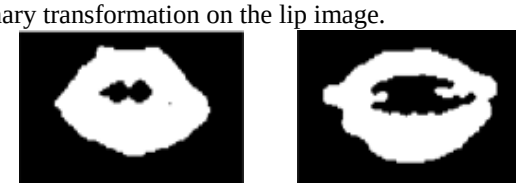

Fig.4 Binary Images.

C. Lip Contour Extraction

Some lip reading system and visual speech recognition systems used lip contour point (10 coordinate points, 14 coordinate points and 16 coordinate points) to extract lip shape. It is not necessary (redundant) to use all or some of the contour points of the lip to define the lip shape where the height and width of the mouth supported by the bounding ellipse are sufficient to approximate the outer lip contour. In this system, twelve coordinate points are taken from the lip boundary contour by splitting the image resulted from lip ROI extraction stage into vertically six pieces to extract lip contour and lip features.

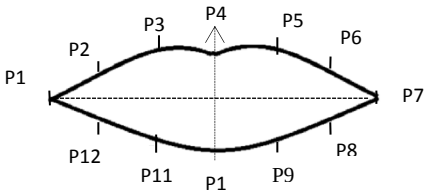

Fig. 5 Twelve key points representing lip movement pattern.

The proposed system adopts 12 coordinate points in order to produce the exterior lip contour, which is more flexible and physically significant compared to the designs based on less points. The shape of the lip can be divided into two parts where the coordinate point set{P1,P2,P3,P4,P5,P6,P7} represents the portion of the upper lip points and{ P1,P12,P11,P10,P9,P8,P7} describes the part of the lower lip points. Even for quite asymmetrical mouths, its high flexibility and variability can model the lip shapes. Fig.5 showed twelve key points representing lip movement pattern. In order to find these twelve coordinate points, Moore Neighborhood Tracing Algorithm is evaluated, in which twelve coordinate points are selected. In the proposed system, we extracted lip contour on lip boundary accurately by using Moore Neighbor Algorithm.

Fig.6 demonstrates some results of lip extraction using the proposed technique. As can be seen, the contour of the lip for distinct mouth shape can be obtained correctly. On a large amount of lip images gathered from various speakers with unadorned lips, we conducted lip extraction, uttering consonants.

Fig. 6 Lip contour extraction results on different lip shape.

IV. Moore Neighborhood Tracing Algorithm

Four of the most popular algorithms for contour tracing, namely: Square Tracing algorithm, Moore-Neighborhood Tracing Algorithm, Radial Sweep Algorithm, Theo Pavlidis' Algorithm. For the proposed system, Moore Neighborhood Tracing algorithm is used for lip contour extraction and movement tracking. A Moore Neighborhood algorithm is used to find outer points on upper and lower lips boundary points. Horizontal and vertical points of the contour are identified after finding lip contour. Horizontal points are lip edges which will be found by maximum extension in the horizontal axis of the contour. Vertical points are the height

of the lips which is used to cover the boundary of the lips and they are maximum extension of vertical axis. Fig.7 shows the contour following sequences of Moore-Neighborhood algorithm.

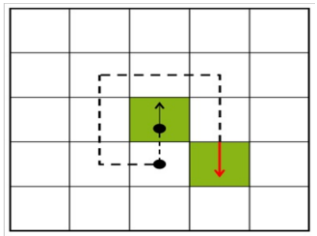

Fig.7 Contour-following sequence of Moore-neighbor tracing (MNT).

Formal description of the Moore-Neighbor tracing algorithm:

Input: A square tessellation, T, containing a connected component P of black cells.

Output: A sequence B (b1, b2 ,..., bk) of boundary pixels i.e. the contour.

Define M(a) to be the Moore Neighborhood of pixel a.

Let p denote the current boundary pixel.

Let c denote the current pixel under consideration i.e. c is in M(p).

```
Begin
    Set B to be empty.
     From bottom to top and left to right scan the cells
     of  T until a black pixel, s, of P is found.
     Insert s in B.
     Set the current boundary point p to s i.e. p=s
     Backtrack i.e. move to the pixel from which s was
     entered.
     Set c to be the next clockwise pixel in M(p).
         While c not equal to s do
             If c is black
                insert c in B
                set p=c
                backtrack (move the current pixel c to
                the pixel from which p was entered)
               else
                 advance the current pixel c to the next
                 clockwise pixel in M(p)
          end While
end
```

V. Feature Extraction

Extraction of features is an essential component of any strategy to lip-reading. Different visual feature extraction methods have been suggested. Extraction of Discrete Cosine Transform (DCT) function was used in[17]. Also used for lip reading were Principle Component Analysis (PCA)[18], Discrete Wavelet Transform (DWT)[18] and Linear Discriminant Analysis (LDA)[17]. Neural networks have been increasingly used in latest years in the classification of images, recognition of images, recognition of expression, lipreading[15],[19],[20], etc. Also used in lip reading[14],[16] were Convolutional Neural Networks (CNNs). CNNs have significantly less links and parameters than traditional neural networks. Training the network is therefore much easier. K. Noda et al. suggested an audio-visual voice recognition CNN of seven layers[14]. This article uses a dynamic function and geometric features to propose a lip-reading technique. An initial lip image sequence, frequently used in studies on lip reading, is substituted by a vibrant image feature. Geometric data was used to extract features.

This paper uses a dynamic features and geometric features to propose a lip-reading strategy. An initial lip image sequence, frequently used in research on lip reading, is substituted by a dynamic image feature. In this paper, for feature extraction, geometrical information was used.

We first properly normalize and rotate the exterior lip contours in the geometric features to compensate for relative location variations. The contour extracts geometric characteristics. The extracted characteristics are the most informative for lip reading, namely lip height (H), width (W), and region (A). These features are extracted based on lip model. Fig. 8 shows the lip model. For this system, we extracted features on only 20 frames for two syllable consonants because of the main difficulties in analyzing of these time series. The facts that their lengths not only differ between the spoken consonants, but also differ between different speakers uttering the single consonant and between the different occasions when the single consonant is uttered by the single speaker.

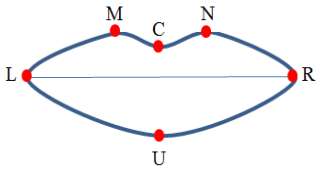

Fig.8 Lip model.

The three features of formulated as the following equations:

$$H = max(M_y, N_y) - U_y \quad (1)$$

$$D = R_x - L_x \quad (2)$$

'A' denotes Area of lip contour,

$$A = \Sigma_x \Sigma_y f(x, y) \quad (3)$$

VI. Experimental Results

The proposed system was evaluated using own audio visual database that consists of different people from different backgrounds. All lip reading system were applied on own audio visual TMC database. The TMC database is composed of 10 subjects, 1 male and 9 female, which have mixture of white and black complexions with no particular lipstick. Each recording include frontal face color video sequences of each speaker, with 14 two syllable Myanmar consonants. Myanmar Consonants composed of 33 consonants. 8 consonants have one syllable, 14 consonants have two syllable, 10 consonants have three syllable, and 1 consonant have four syllable.

These database was captured in three lighting system to control illumination conditions by Sony DVCan-DSR 300A professional video camera with FUJINON TVZOOM LENS. Videos are recorded in mp4 format with frame rate of 30 fps at a resolution of 720×480 pixels. Speaker's utterances were recorded on one time. Table.1 shows the description about database participants. Fig. 9 shows more experimental result examples. It can be clearly observed that the lip contours can be accurately extracted.

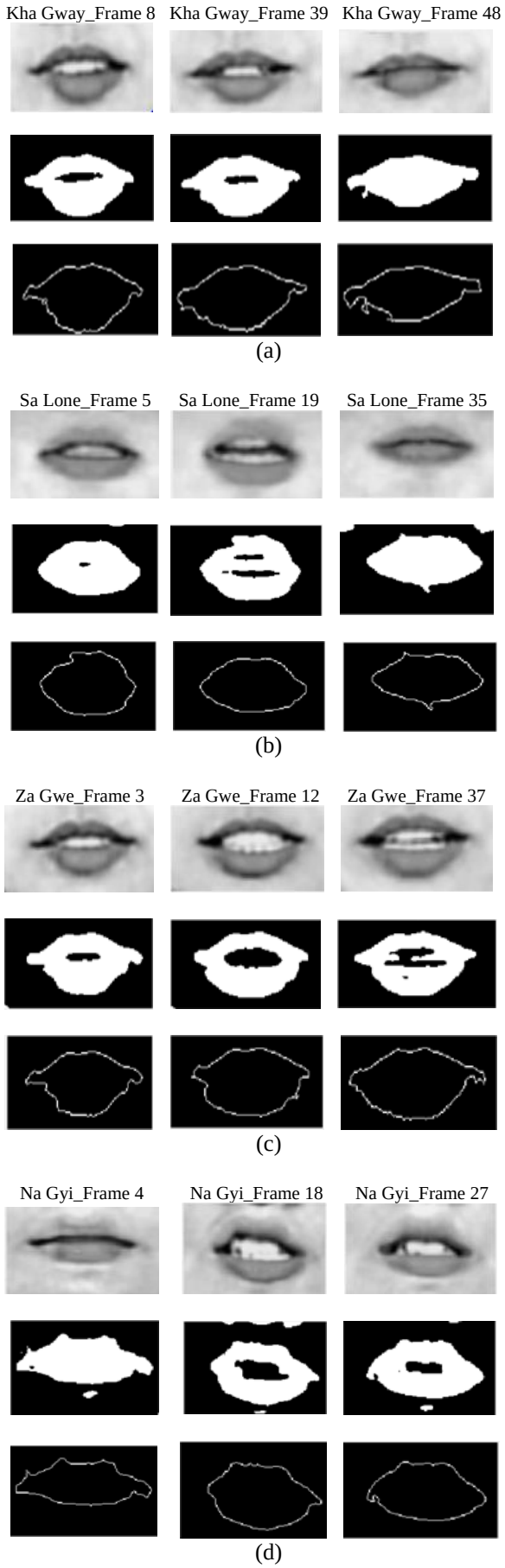

Fig.9 First row of (a), (b), (c) and (d) show the results of transformed images of L channel, second row of (a), (b), (c) and (d) show the binary images, third row of (a), (b), (c) and (d) show contour extracted on different frames on different consonants on different lip shapes.

TABLE.1 DESCRIPTION ABOUT DATABASE PARTICIPANTS

Subject	Sex	Age	Job
Speaker 1	F	29	Student
Speaker 2	F	28	Student
Speaker 3	M	30	Student
Speaker 4	F	10	Student
Speaker 5	F	34	Student
Speaker 6	F	30	Student
Speaker 7	F	33	Student
Speaker 8	F	30	Student
Speaker 9	F	33	Student
Speaker 10	F	36	Student

We need to detect and extract lip portion for these separated frames. The task is performed by color based object detection method.

It is hard to extract a number of distinctive characteristics from the visual signal to represent a consonant because distinct individuals are speaking in distinct ways, creating a range of visual signals for the single consonant. The facts that their lengths not only differ between the spoken consonants, but also between separate speakers speaking the single consonant and the distinct times when the single consonant is spoken by the single speaker. Hence 15 significant lip portion frames are selected for each consonants utterance to extract features. Table 2 shows the accuracy on proposed features and this table also show the training time and testing time for one syllable consonants and two syllable consonants.

TABLE.2 CLASSIFICATION ACCURACY ON PROPOSED FEATURES

Features	Accuracy %		Training Time		Testing Time	
	One sylla-ble	Two sylla-ble	One sylla-ble	Two sylla-ble	One sylla-ble	Two sylla-ble
Area, Width, Height	**81.3**	**93.2**	4.55	13.5	14.1	66.1

VII. CONCLUSION

In this paper, we implemented the lip event recognition and extracted lip' movement features for 14 Myanmar consonants: (□ (Ka Gyi)□ □ (Kha Gway) □ □ (Ga Nge) □ □ (Ga Gyi) □ □ (Sa Lone) □ □ (Sa Lain) □ □ (Za Gwe) □ □ (Da Dway) □ □ (Na Gyi) □ □ (Na Nge) □ □ (Pa Saug) □ □ (Ba Gone) □ □ (Ya Gaug) □ □ (La Gyi)). The lip movement detection and recognition results of the proposed method give acceptable and significant results for feature extraction process. The proposed system uses geometric information and a manual selection of a pixel point is required for initializing the lip contour detection. As a result of analysis on features, □ (Sa Lone), □ (Da Dway); the mouth shall be opened to the maximum, the height shall be increased. Other syllable consonants □ (Da Dway), □ (Pa Sauk), □ (Ba Gone) are the width expands to the maximum. The experimental result achieves significant accuracy and training speed is fast. For feature work, we intended to lip reading system for the remaining Myanmar consonants and other language such as English, Japanese with Myanmar speakers.

REFERENCES

[1] E. Skodras and N. Fakotakis, "An unconstrained method for lip detection in color images," IEEE, Acoust., Speech, Signal Process., Prague, Czech Republic, May 2011, pp. 1013–1016.
[2] P. Delmas, N. Eveno, and M. Liévin, "Towards robust lip tracking," IEEE , Pattern Recognit., Quebec City, Canada, 2002, pp. 528–531.
[3] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic lip tracking," IEEE Trans. Circuits System. Video Technol., vol. 14, no. 5, pp. 706–715, May 2004.
[4] A. Nefi an, L. Liang, X. Pi, L. Xiaoxiang, C. Mao and K. Murphy, "A couple HMM for Audio-Visual Speech Recognition" ICASSP', 02, 2002, pp. 2013-2016.
[5] P. Delmas, P.-Y. Coulon and V. Fristot, "Automatic Snakes For Robust Lip Boundaries Extraction", ICASSP'99, 1999, pp. 3069-3072.
[6] M. Kass, A. Witkin, D. Terzopoulos. "Snakes: Active contour models," Int. Journal of Computer Vision, 1(4), pp 321-331, jan. 1988.
[7] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," IEEE Trans Pattern Analysis and Machine Intelligence, 15(6), pp. 569-579, June 1993.
[8] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-Visual Speech Recognition Using MPEG-4 Compliant Visual Features," EURASIP Journal on Applied Signal Processing, Spec. Issue on Joint Audio-Visual Speech Processing, pp. 1213-1227, Sept. 2002.
[9] S.L. Wang , A.W.C. Liew , W.H. Lau , S.H. Leung , "An automatic lipreading system for spoken digits with limited training data, IEEE Trans. Circuits Syst. Video Technol," 18 (12) (2008) 1760–1765 .
[10] M. Faraj , J. Bigun, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," IEEE Trans. Comput. 56 (9) (2007) 1169–1175.
[11] M. Bendris , D. Charlet , G. Chollet , "Lip activity detection for talking faces classification," Proceedings of International Conference on Machine Vision, 2010, pp. 187–190.
[12] Y. Tian , T. Kanade , J. Cohn , "Recognizing action units for facial expression analysis", IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.
[13] Lai Pei Mei. "Interpretation Of Alphabets By Images Of Lips Movement For Native Language", University of Technologies, Malaysia, 2014.
[14] K. Noda, Y. Yamaguchi, K. Nakadai, et al., "Audio-visual speech recognition using deep learning," Applied Intelligence, pp. 722-737, 42(4), 2015.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification
[16] with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097-1105, 2012.
[17] Y. Takashima, Y. Kakihara, R. Aihara, et al., "Audio-Visual Speech Recognition Using Convolutive Bottleneck Networks for a Person with Severe Hearing Loss," IPSJ Transactions on Computer Vision and Applications, pp. 64-68, 7(0), 2015.
[18] X. Hong, H. Yao, Y. Wan, et al., "A PCA based visual DCT feature extraction method for lip-reading," Intelligent Information Hiding and Multimedia Signal Processing, pp. 321-326, 2006.
[19] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," ICIP, pp. 173-177, 1998.
[20] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: integrating automatic speech recognition and lip-reading," ICSLP, pp. 547-550, 1994.
[21] A. Bagai, H. Gandhi, R. Goyal, et al., "Lip-reading using neural networks [J]. International Journal of Computer Science and Network, 9(4), pp. 108-111, 2009.
[22] Y. Tian, T. Kanade and J. Cohn, "Robust Lip Tracking by Combining Shape, Color and Motion", In Proc. ACCV'00, 2000.
[23] T. Coianiz, L. Torresani and B. Caprile, "2D Deformable Models for Visual Speech Analysis.", In NATO Advanced Study Institute: Speech reading by Man and Machine, 1995, pp. 391-398.
[24] M.E. Hennecke, K.V. Prasad and D.G. Stork, "Using deformable templates to infer visual speech dynamics", In Proc. 28th Annual Asilomar Conference on Signals, Systems, and Computers, 1994, pp. 578-582.
[25] N. Tsapatsoulis, Y. Avrithis and S. Kollias, "Effi cient Face Detection for Multimedia Applications", In Proc. ICIP'00, 2000.
[26] M.H. Yang and N. Ahuja, "Gaussian Mixture Model for Human Skin Color and Its Application in Image and Video Databases", In Proc. of the SPIE : Conf. on Storage and Retrieval for Image and Video Databases, vol. 3656, pp. 458-466, 1999.