Real time facial expression recognition from image sequences using Support Vector Machines

I. Kotsia [a] and I. Pitas[a]

[a]Aristotle University of Thessaloniki,
Department of Informatics,
Box 451, 54124
Thessaloniki, Greece

ABSTRACT

In this paper, a real-time method is proposed as a solution to the problem of facial expression classification in video sequences. The user manually places some of the Candide grid nodes to the face depicted at the first frame. The grid adaptation system, based on deformable models, tracks the entire Candide grid as the facial expression evolves through time, thus producing a grid that corresponds to the greatest intensity of the facial expression, as shown at the last frame. Certain points that are involved into creating the Facial Action Units movements are selected. Their geometrical displacement information, defined as the coordinates' difference between the last and the first frame, is extracted to be the input to a six class Support Vector Machine system. The output of the system is the facial expression recognized. The proposed real-time system, recognizes the 6 basic facial expressions with an approximately 98% accuracy.

Keywords: Facial expression recognition, Facial Action Unit, Facial Action Coding System, Support Vector Machines, Candide grid.

1. INTRODUCTION

During the past two decades, facial expression recognition has attracted a significant interest in the scientific community, due to its importance for human centered interfaces. The facial expressions under examination were defined as a set of six basic expressions (anger, disgust, fear, happiness, sadness and surprise), whose combinations produce every "other" facial expression,[1,2] In order to make the recognition procedure more standardized, a set of muscle movements, known as Action Units, was created by psychologists, thus forming the so called Facial Action Coding System (FACS),[3] .[4] These Action Units are combined into facial expressions according to the rules proposed in.[5]

A survey on automatic facial expression recognition can be found in.[6] Gabor wavelets have been thoroughly used in facial expression recognition. They were used as a preprocessing step in order to create the input for multi-layer neural networks.[7] Gabor wavelets were also combined with elastic graph matching techniques.[8] Facial expression recognition has also been investigated using Tree-Augmented-Naive Bayes (TAN) trees in.[9]

In this paper, a novel real time method for facial expression recognition using Support Vector Machines is proposed. The system is composed of a tracking subsystem, used for information extraction, followed by a Support Vector Machines (SVM) subsystem, used for facial expression classification. The user initially places manually some of the points of the Candide[10] grid on the face depicted at the first frame of the image sequence. The software tracks the Candide grid through time, resulting in a grid that represents the facial expression with the greatest intensity, as it appears at the last frame of the image sequence. From all of the nodes creating the Candide grid, many of them do not play a crucial role when it comes to defining the appearance of facial

Further author information: (Send correspondence to Ioannis Pitas.)
I. Pitas: E-mail: pitas@aiia.csd.auth.gr, Telephone: +302310996304

expressions. Therefore only those that correspond to the Facial Action Units should be taken into consideration for the classification procedure, in order to increase its speed. Thus a subset of the initial points, 62 in number, is chosen, due to their importance in recognizing facial expressions. The geometrical displacement of those points, defined as the difference of each node's coordinates between the first and the last frame of the image sequence, are used as an input to a multiclass SVM system. Each classification class represents one of the 6 basic facial expressions. The experiments were performed using the Cohn-Kanade database and the results show that the above mentioned novel real-time system can achieve an accuracy of 97.75% when recognizing 6 basic facial expressions.

2. SYSTEM DESCRIPTION

The diagram of the system used for the experiments is shown in Figure 1. The system is composed of two

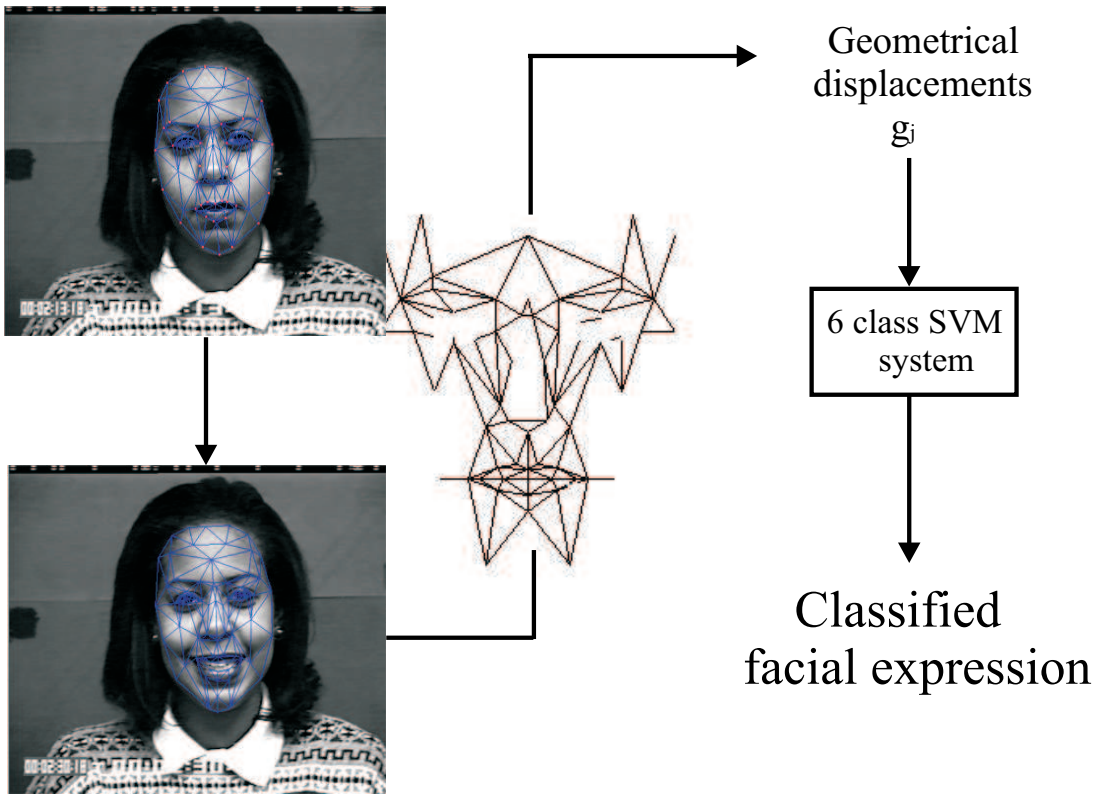

Figure 1. System description

subsystems, one for geometrical information extraction and one for geometrical information classification.

Facial expressions can be described as combinations of Facial Action Units (FAUs), as proposed by.[5] As can be seen from the rules proposed, the FAUs that are necessary for fully describing all facial expressions are the FAUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 16, 17, 20, 23, 24, 25 and 26. Therefore, these 17 FAUs are responsible for creating movement according to the Facial Action Coding System (FACS).

At the second column of Table 1, the rules that define the combinations of FAUS that construct the facial expressions, are shown, according to.[5] From all of these FAUs appearing in the facial expression description rules, many describe two or more facial expressions. For example, FAU 26 appears in every facial expression, thus making its presence impossible to define a facial expression. On the other hand, FAU 9 appears only in facial expression disgust, defining that way uniquely the specific facial expression. Therefore a subset of FAUs

Table 1. The FAUS chosen for facial expression recognition.

Expression	FAUs coded description[5]	FAUs chosen
Anger	$4+7+(((23 \text{ or } 24) \text{ with or not } 17) \text{or } (16+(25 \text{ or } 26))$ or $(10+16+(25 \text{ or } 26)))$with or not 2	23 or 24
Disgust	$((10 \text{ with or not } 17) \text{ or } (9 \text{ with or not } 17))+(25 \text{ or } 26)$	9
Fear	$(1+4)+(5+7)+20+(25 \text{ or } 26)$	20
Happiness	$6+12+16+(25 \text{ or } 26)$	$12+16$
Sadness	$1+4+(6 \text{ or } 7)+15+17+(25 \text{ or } 26)$	15
Surprise	$(1+2)+(5 \text{ without } 7)+26$	5

is chosen (FAUs 5, 9, 12, 15, 16, 20, 23 and 24) as those that appear once or twice in the whole set of facial expressions. Those FAUs can accurately describe the 6 basic facial expressions with an accuracy equal to 94% according to the rules proposed in.[5] The FAUs chosen the way described above are shown at the third column of Table 1.

2.1. Geometrical displacement information extraction

The tracking subsystem performs grid node information extraction by a grid adaptation system, based on deformable grid models.[11] Facial wireframe node tracking is performed by a pyramidal variant of the well-known Kanade-Lucas-Tomasi (KLT) tracker.[12] The loss of tracked features is handled through a model deformation procedure that increases the robustness of the tracking algorithm. Tracking initialization is performed in a semi-automatic fashion, i.e., the facial wireframe model is fitted to an image representing a neutral facial expression, exploiting physics-based deformable shape modeling. The Candide grid is randomly initialized on the first frame of the image sequence, being in its neutral state. The user has to manually place some of its nodes to the face depicted. The nodes around the eyes, eyebrows and mouth are the ones with the greatest importance, since they are the ones responsible for the formation of movement according to FACS. The software automatically adjusts the grid to the face and then tracks it through the image sequence, following the facial expression evolving through time.[13] At the end, the grid adaptation software produces the deformed Candide grid that corresponds to the facial expression appearing at the image sequence.

The deformed Candide grid produced by the grid adaptation software, that corresponds to the greatest intensity of the facial expression shown, is constructed by 104 nodes. Not all of these nodes are important for the recognition of the facial expression, for example the contour of the face doesn't affect the way the eyes or the mouth nodes move. Thus, a subset of 62 nodes are chosen, as those that control the movement described by the 17 FAUs used for describing facial expressions. The grid that consists of these nodes can be seen in Figure 1.

The classification is performed based only in geometrical information, without taking into consideration any luminance or color information. The geometrical displacement information of the subset's nodes coordinates is extracted to be used for facial expression classification.

Let $\mathcal{U}$ the database that contains the geometrical displacement information separated into the 6 different classes, $\mathcal{U}_k(\, k \in \{1,\ldots,6\})$, each one representing one of six basic facial expressions (anger, disgust, fear, happiness, sadness and surprise).

The geometrical information used is the displacement of one node $\mathbf{d}_j^i$, defined as the difference between the last and the first frame's coordinates

$$\mathbf{d}_j^i = \begin{bmatrix} \Delta x_j^i \\ \Delta y_j^i \end{bmatrix}, \quad i \in \{1,\ldots,K\} \quad \text{and} \quad j \in \{1,\ldots,N\} \tag{1}$$

where i is the number of nodes taken under consideration, here K, equal to 62 and j is the the number of image sequences to be examined, here N, equal to 222.

In that way, for every image sequence to be examined, a feature vector $\mathbf{g}_j$ that belongs to one of the six facial expression classes $\mathcal{U}_k$ is constructed, containing the geometrical displacement of every node taken into consideration, thus having the following form

$$\mathbf{g}_j = \begin{bmatrix} \mathbf{d}_j^1 \\ \mathbf{d}_j^2 \\ \vdots \\ \mathbf{d}_j^K \end{bmatrix}. \tag{2}$$

where the vector $\mathbf{g}_j$ has $F = 62 \cdot 2 = 124$ dimensions.

2.2. Geometrical displacement information classification

The feature vector $\mathbf{g}_j \in \Re^F$ is used as an input to a multi class Support Vector Machine system. Six classes were considered for the experiments, each one representing one of the basic facial expressions (anger, disgust, fear, happiness, sadness and surprise). The SVM system, classifies each set of the grid's geometrical displacements to one of the six basic facial expressions.

More specifically, as an input for the SVM system, the feature vector $\mathbf{g}_j$ is used, labelled properly with the true corresponding facial expression. The output of the SVM system is a label that classifies the grid under examination to one of the six basic facial expressions.

3. SUPPORT VECTOR MACHINES

A test $\mathbf{g}_j$ displacement vector has to be classified to one of the six facial expressions. This is done using multiclass SVMs[14] that are a generalization of the binary SVM.[15]

The SVMs creates a decision function $f(\mathbf{g}_j, \boldsymbol{\alpha})$ which classifies a vector $\mathbf{g}_j$ into one of the six basic facial expression classes. The vector $\boldsymbol{\alpha}$ should be chosen in such a way that for any $\mathbf{g}_j$ the function should be able to provide a classification $l_j \in \{1 \dots 6\}$ (class label).

The main idea of an SVM system is to construct a hyperplane that will separate the desired classes, in such a way that the margin (defined as the distance between the hyperplane and the nearest node) is maximal. Therefore, to generalize, the following equation should be minimized

$$\Phi(\mathbf{w}, \xi) = 1/2 \sum_{m=1}^{6} ({\mathbf{w}_m}^T \cdot \mathbf{w}_m) + c \cdot \sum_{i=1}^{N} \sum_{m \neq l_i} \xi_i^m \tag{3}$$

with constraints

$$({\mathbf{w}_{l_i}}^T \cdot \mathbf{g}_i) + b_{l_i} \geq ({\mathbf{w}_m}^T \cdot \mathbf{g}_i) + b_m + 2 - \xi_i^m \tag{4}$$

$$\xi_i^m \geq 0, \quad i \in \{1, \dots, N\} \quad m \in \{1, ..., 6\} \backslash l_i. \tag{5}$$

The decision function that is derived from equation 3 is the following

$$f(\mathbf{g}_j) = \arg\max_n [({\mathbf{w}_n}^T \cdot \mathbf{g}_j) + b_n], \quad n \in \{1, \dots 6\} \tag{6}$$

The solution to this optimization problem in dual variables can be found by the saddle node of the Lagrangian

$$L(\mathbf{w},\mathbf{b},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = 1/2\sum_{m=1}^{6}({\mathbf{w}_m}^T\cdot\mathbf{w}_m) + c\sum_{i=1}^{N}\sum_{m=1}^{6}\xi_i^m$$
$$-\sum_{i=1}^{N}\sum_{m=1}^{6}\alpha_i^m[((\mathbf{w}_i-\mathbf{w}_m)^T\cdot\mathbf{g}_i)+b_{l_i}-b_m-2+\xi_i^m]$$
$$-\sum_{i=1}^{N}\sum_{m=1}^{6}\beta_i^m\xi_i^m \tag{7}$$

with the variables

$$\alpha_i^{l_i}=0,\quad \xi_i^{l_i}=2,\quad \beta_i^{l_i}=0,\quad i=\{1,\ldots,N\} \tag{8}$$

and constraints

$$\alpha_i^m\geq 0,\quad \beta_i^m=0,\quad \xi_i^m\geq 0, \tag{9}$$
$$i\in\{1,\ldots,N\}\quad m\in\{1,...,6\}\backslash l_i \tag{10}$$

which has to be maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and be minimized with respect to $\mathbf{w}$ and $\boldsymbol{\xi}$.

By further processing[14] equation (6) is finally expressed as

$$f(\mathbf{g}_j,\alpha)=\arg\max_n[\sum_{i:l_i=n}A_i({\mathbf{g}_i}^T\cdot\mathbf{g}_j)-\sum_{i:l_i\neq n}\alpha_i^n({\mathbf{g}_i}^T\cdot\mathbf{g}_j)+b_n] \tag{11}$$

where $\boldsymbol{\alpha}$ is the vector of Lagragian multipliers in equation (7) and A_i is defined as

$$A_i=\sum_{m=1}^{6}\alpha_i^m. \tag{12}$$

The previous analysis is used for linear decision surfaces. For the proposed method, nonlinear SVMs were considered. To achieve that, a nonlinear mapping to a high dimensional feature mapping, $Z(\mathbf{g}_j)$ was used. This mapping is defined by a positive kernel function, $k({\mathbf{g}_j}^T,\mathbf{g}_j)$, specifying an inner product in the feature space

$$Z({\mathbf{g}_j}^T)\cdot Z(\mathbf{g}_j)=k({\mathbf{g}_j}^T,\mathbf{g}_j). \tag{13}$$

The kernel used for the experiments was a d degree polynomial function, defined in general as

$$k({\mathbf{g}_j}^T,\mathbf{g}_j)=({\mathbf{g}_j}^T\cdot\mathbf{g}_j+1)^d. \tag{14}$$

4. EXPERIMENTAL RESULTS

The database used for the experiments was the Cohn-Kanade database,[4] which is encoded into combinations of Action Units. These combinations were translated into facial expressions according to.[5] For each person, the image sequence was created and processed by the grid adaptation system, based on deformable models. In figure 2, a sample of image sequences of one person from the database used for the experiments, is shown.

The database created for the experiments was of limited size, therefore the classifier accuracy was measured using the leave-one out method[16] in order to make maximal use of the available data and produce averaged accuracy results. All image sequences were divided into 6 classes, each one corresponding to one of the 6 basic facial expressions to be recognized. Each class consisted of the same number of image sequences, here equal to

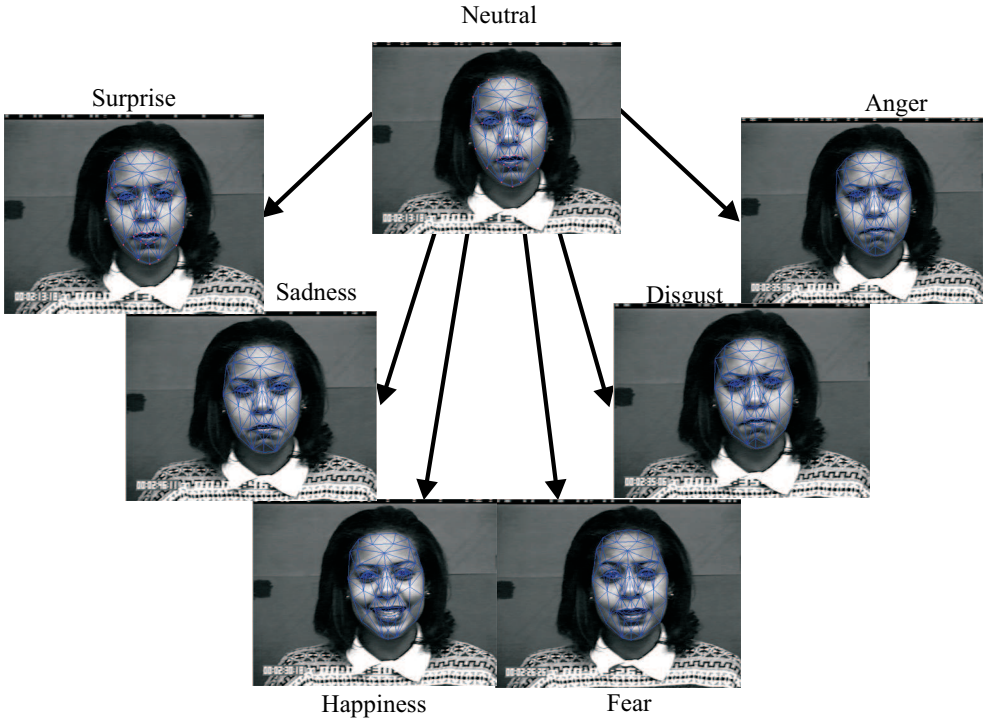

Figure 2. Example of a poser for all 6 basic facial expressions

37. The training set was formed by the 36 image sequences of each facial expression, while the remaining one image sequence for each facial expression was used to form the testing set. The recognition was performed and then the first image sequence for each facial expression was used to form the new testing set, while the previous testing set was incorporated in the new training set. This was performed until all of the image sequences were used as testing sets. The classification accuracy was measured as mean value of the percentages of the correctly classified facial expressions. The polynomial function used for the creation of the polynomial kernel, was of degree 3.

The experiments indicated that the whole system is fast enough to fulfill a real-time system's requirements, since it is able to classify 20 frames per second. The accuracy achieved was equal to 97,75% when the 6 basic facial expressions were under examination.

The accuracy obtained is averaged over all facial expressions and does not provide any information with respect to a particular expression. The confusion matrix [17] has been computed to handle this problem. It is a $n \times n$ matrix containing the information about the actual class label l_j, $j = 1, .., n$ (in its rows) and the label obtained through classification p_j, $j = 1, .., n$ ones (in its columns). The diagonal entries of the confusion matrix are the number of facial expressions that are correctly classified, while the off-diagonal entries correspond to misclassification. The confusion matrix showed that the ambiguous facial expression was anger, since it was the only one misclassified as another one of the remaining 5 basic facial expressions. More specifically, anger was mostly misclassified as sadness and then as disgust, as shown from the confusion matrix 2.

The abbreviations an, di, fe, ha, sa and su represent anger, disgust, fear, happiness, sadness and surprise respectively, while lab_{ac}, lab_{clas} represent the actual and the classified label of the video sequence, respectively.

Table 2. Confusion matrix for the 6 basic facial expressions

$lab_{ac} \backslash lab_{clas}$	an	di	fe	ha	sa	su
an	32	0	0	0	0	0
di	1	37	0	0	0	0
fe	0	0	37	0	0	0
ha	0	0	0	37	0	0
sa	4	0	0	0	37	0
su	0	0	0	0	0	37

5. CONCLUSION

Facial expression recognition using Support Vector Machines has been investigated in this paper. The user adjusts the Candide grid to the face depicted at the first frame of the image sequence, by manually placing some of its nodes. The grid adaptation system, based on deformable models, tracks the grid, as the facial expression progresses through the time, thus producing a deformed grid that corresponds to the highest intensity of the facial expression under examination. A subset of the deformed grid's nodes is chosen, as those that are the most important for the Facial Action Units formation. Their geometrical displacement information, defined as the difference between the first and the last frame, is used as an input to a six class (one for each facial expression) Support Vector Machine System. The output of the Support Vector is the facial expression recognized from the image sequence. The above mentioned novel real-time system achieves a recognition rate of approximately 98%, which is the best achieved, according to the authors knowledge of the facial expression recognition literature.

ACKNOWLEDGMENTS

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding Through Semantic Computation and Learning" (FP6-507752)

REFERENCES

1. P. Ekman and W. Friesen, *Emotion in the Human Face*, Prentice Hall, New Jersey, 1975.
2. P. Ekman, *Unmasking the Face*, Cambridge University Press, Cambridge, 1982.
3. P. Ekman and W. Friesen, "Manual for the facial action coding system," *Consulting Psychologists Press* , 1977.
4. J. C. T. Kanade and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of IEEE International Conference on Face and Gesture Recognition*, pp. 46–53, March 2000.
5. M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, pp. 1424–1445, December 2000.
6. M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing* **18**, pp. 881–905, August 2000.
7. W. Fellenz, J. Taylor, N. Tsapatsoulis, , and S. Kollias, "Comparing template-based, feature-based and supervised classification of facial expressions from static images," *Computational Intelligence and Applications, World Scientific and Engineering Society Press* , 1999.
8. L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 1997.
9. A. M. I.Cohen, N.Sebe and T.S.Huang, "Facial expression recognition from video sequences," in *Proceedings of IEEE International Conference on International Conference on Multimidia &Expo*, 2002.
10. M. Rydfalk, "Candide: A parameterized face," tech. rep., Linkoping University, 1978.

11. S. Krinidis and I. Pitas, "2-D physics-based deformable shape models: Explicit governing equations," in *Proceedings of First International Workshop on "Interactive Rich Media Content Production: Architectures, Technologies, Applications, Tools"*, pp. 43–55, (Lausanne, Switzerland), 16-17 October 2003.
12. J. Y. Bouguet, "Pyramidal implementation of the Lucas-Kanade feature tracker," tech. rep., Intel Corporation, Microprocessor Research Labs, 1999.
13. S. Krinidis and I. Pitas, "Statistical analysis of facial expressions for facial expression synthesis.," *submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence* , 2004.
14. J.Weston and C.Watkins, "Multi-class support vector machines," Tech. Rep. Techical report CSD-TR-98-04, 2004.
15. V. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, 1995.
16. T. M. Cover, *Learning in pattern recognition*, Academic Press.
17. M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **21**(12), pp. 1357–1362, 1999.