INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

XBeats-An Emotion Based Music Player

Sayali Chavan[1], Ekta Malkan[2], Dipali Bhatt[3], Prakash H. Paranjape[4]

[1]U.G. Student, Dept. of Computer Engineering,
D.J. Sanghvi College of Engineering,
Vile Parle (W), Mumbai-400056.
sayalichavan17@gmail.com

[2] U.G. Student, Dept. of Computer Engineering,
D.J. Sanghvi College of Engineering,
Vile Parle (W), Mumbai-400056.
ekta.malkan27@yahoo.in

[3] U.G. Student, Dept. of Computer Engineering,
D.J. Sanghvi College of Engineering,
Vile Parle (W), Mumbai-400056.
dipupb1392@gmail.com

[4] Assistant Professor, Dept. of Computer Engineering,
D.J. Sanghvi College of Engineering,
Vile Parle (W), Mumbai-400056.
prakashparanjape2012@gmail.com

Abstract*: Music expresses emotion. Music's interconnection with society can be seen throughout history. Music seems to be one of the basic actions of humans. Music is ubiquitous in our daily life. People actively or passively listen to music and consciously or non-consciously experience it as a form of emotion expression. In this paper, we present a new emotion aware and user-interactive music system, XBeats. It aims to provide the user-preferred music with emotion awareness. The system starts recommendation with expert knowledge. If the user does not like the recommendation, he/she can decline the recommendation and select the desired music himself/herself. Thus, this player focuses on automating and enhancing the process of listening to music.*

Keywords: *Emotion, Expression Recognition, Face detection, Feature extraction, Music.*

1. INTRODUCTION

Face plays significant role in social communication. Interpreting non-verbal face gestures is used in a wide range of applications. An intelligent user-interface not only should interpret the face movements but also should interpret the user's emotional state (Breazeal, 2002). Knowing the emotional state of the user makes machines communicate and interact with humans in a natural way: intelligent entertaining systems for kids, interactive computers, intelligent sensors, social robots, etc to name a few. An emotion based music player is one such Human-computer interaction device. It detects the face of the user by capturing an image using webcam, recognizes his expressions, identifies the emotional state of the user and plays music best suited to the user's emotional state. The system is built in traditional manner and consists of 4 stages: face detection and tracking, feature extraction, emotion recognition and playing music.

2. LITERATURE REVIEW

2.1 Face detection methods

The techniques for face detection can be distinguished into two groups: holistic, where face is treated as a whole unit and analytic, where co-occurrence of characteristic facial elements is studied.

Pantic and Rothkrantz [15] proposed system which process images of frontal and profile face view. Vertical and horizontal histogram analysis is used to find face boundaries. Then, face contour is obtained by thresholding the image with HSV color space values.

Kobayashi and Hara [5] used image captured in monochrome mode to find face brightness distribution. Position of face is estimated by iris localization.

2.2 Feature extraction methods

Pantic and Rothkrantz [15] selected a set of facial points from frontal and profile face images. The expression is measured by a distance between position of those points in the initial image (neutral face) and peak image (affected face).

Cohn et al. [16] developed geometric feature based system in which the optical flow algorithm is performed only in 13x13 pixel regions surrounding facial landmarks.

Shan et al. [3] investigated the Local Binary Pattern method for texture encoding in facial expression description. Two methods of feature extraction were proposed. In the first one, features are extracted from fixed set of patches and in the second method from most probable patches found by boosting.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

2.3 Expression Recognition

The last part of the FER system is based on machine learning theory; precisely it is the classification task. The input to the classifier is a set of features which were retrieved from face region in the previous stage. Classification requires supervised training, so the training set should consist of labeled data. There are a lot of different machine learning techniques for classification task, namely: K-Nearest Neighbors, Artificial Neural Networks, Support Vector Machines, Hidden Markov Models, Expert Systems with rule based classifier, Bayesian Networks or Boosting Techniques (Adaboost, Gentleboost).

Major problems encountered in the above stages are different scales and orientations of face. They are usually caused by subject movements or changes in distance from camera. Significant body movements can also cause drastic changes in position of face in consecutive frames what makes tracking harder. Complexity of background and variety of lightning conditions can be also quite confusing in tracking. For instance, when there is more than one face in the image, system should be able to distinguish which one is being tracked. Last but not least, occlusions which usually appear in spontaneous reactions need to be handled as well.

3. CURRENT MUSIC SYSTEMS

The features available in the existing Music players present in computer systems are as follows:

- Manual selection of Songs
- Party Shuffle
- Playlists
- Music squares where user has to classify the songs manually according to particular emotions for only four basic emotions .Those are Passionate, Calm, Joyful and Excitement.

Limitations of existing system:

- It requires the user to manually select the songs.
- Randomly played songs may not match to the mood of the user.
- User has to classify the songs into various emotions and then for playing the songs user has to manually select a particular emotion.

4. PROPOSED SYSTEM

In this paper, we consider the notion of collecting human emotion from the user's expressions, and explore how this information could be used to improve the user experience with music players. We present a new emotion based and user-interactive music system, XBeats. It aims to provide user-preferred music with emotion awareness. The system starts recommendation with expert knowledge. If the user does not like the recommendation, he/she can decline the recommendation and select the desired music himself/herself. This paper is based on the idea of automating much of the interaction between the music player and its user. It introduces a "smart" music player, XBeats that learns its user's emotions, and tailors its music selections accordingly. After an initial training period, XBeats is able to use its internal algorithms to make an educated selection of the song that would best fit its user's emotion.

4.1 Face detection and tracking

First part of the system is a module for face detection and landmark localization in the image. The algorithm for face detection is based on work by Viola and Jones [1]. In this approach image is represented by a set of Haar-like features. Possible types of features are two-, three- and four rectangular features (Fig. 1).

Feature value is calculated by subtracting sum of the pixels covered by white rectangle from sum of pixels under gray rectangle. Two rectangular features detect contrast between two vertically or horizontally adjacent regions. Three rectangular features detect contrasted region placed between two similar regions and four rectangular features detect similar regions placed diagonally (Fig. 3).

Rectangle features can be computed very rapidly using an intermediate representation for the image which we call the integral image. The integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive:

$$ii(x, y) = \sum_{x' \le x, y' \le y} i(x', y')$$

where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image. Using the following pair of recurrences:

$$s(x, y) = s(x, y - 1) + i(x, y) \quad (1)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y) \quad (2)$$

where $s(x, y)$ is the cumulative row sum, $s(x,-1) = 0$, and $ii(-1, y) = 0$, the integral image can be computed in one pass over the original image. Thus feature can be computed rapidly because the value of each rectangle requires only 4 pixel references (Fig. 2).

Having the representation of the image in rectangular features, the classifier needs to be trained to decide if the image contains searched object (face) or not. The number of features is much higher than the number of pixels in the original image.

However, it was proven that even a small set of well-chosen features can build a strong classifier. That is why, the AdaBoost algorithm can be used for training. Each step selects the most discriminative feature which separates positive and negative examples in the best way. The method is widely used in area of face detection. However, it can be trained to detect any object. This algorithm is quick and efficient and could be used in real-time applications.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

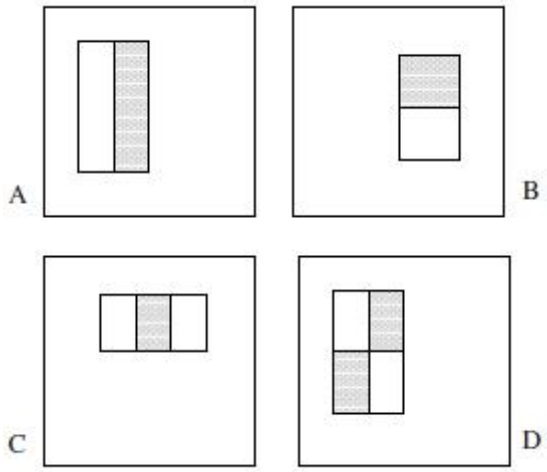

Figure 1: Examples of Haar-like features [1]. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Two-rectangle features are shown in (A) and (B). Figure (C) shows a three-rectangle feature, and (D) a four-rectangle feature.

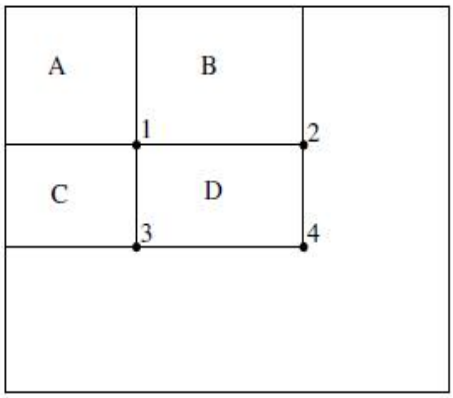

Figure 2: Integral image concept [1]. The sum of the pixels within rectangle *D* can be computed with four array references. The value of the integral image at location 1 is the sum of the pixels in rectangle *A*. The value at location 2 is *A* + *B*, at location 3 is *A* + *C*, and at location 4 is *A* + *B* + *C* + *D*.
The sum within *D* can be computed as 4 + 1 − (2 + 3).

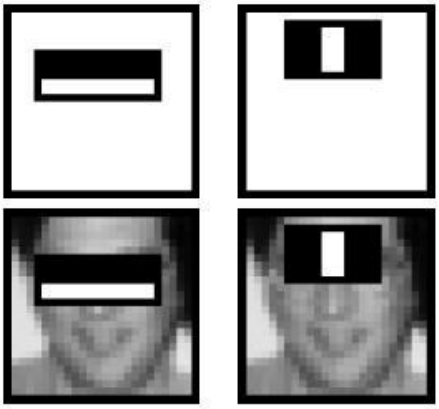

Figure 3: Haar-like features detected on a face [1]

As seen in Figure 3, the first and second features are selected by AdaBoost. The two features are shown in the top row and then overlaid on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose.
The face detection procedure includes some steps which are consecutively performed on the input image. Firstly, the classifier trained for face detection searches for a face in the image (Fig. 4). In case when face is not found in the image, further processing is omitted and system returns appropriate error message.

Figure 4: Face detection procedure

If the face is located, the classifiers for eye detections are employed only on the upper part of the face. The left and right eyes are detected separately – in left and right upper face regions (Fig. 5). Finally, the mouth region is located with the fourth classifier which searches in the lower part of the face. The search area of facial elements detectors is narrowed in order to improve the time efficiency of the algorithm.

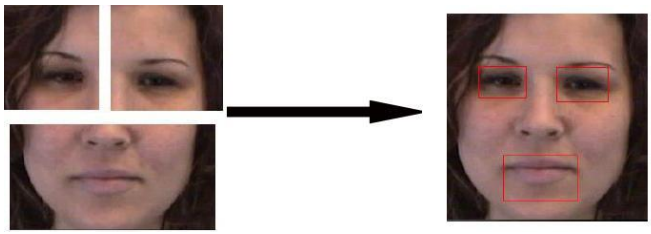

Figure 5: Face elements localization

Having locations of the face and facial landmarks, the face representation can be formed. If there are more faces detected in the image, the algorithm takes the biggest one for further processing.

4.2 Feature extraction

Facial feature representation is to derive a set of features from original face images which minimizes within class variations of expressions whilst maximizes between class variations. There are two main types of approaches to extract facial features: geometric feature-based methods and appearance-based methods [6]. Gabor-wavelet appearance features were demonstrated to be more effective than geometric features [7], and work better in real-world environments [8]. Although Gabor-wavelet

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

representations have been widely adopted [7, 9, 8], it is computationally expensive to convolve face images with multi-banks of Gabor filters in order to extract multi-scale and orientational co-efficients. Due to this fact, another method called Local Binary Patterns (LBP) gains more popularity in facial texture analysis.

The original LBP operator was introduced by Ojala et al [2]. The operator labels the pixels of an image by thresholding the 3×3 neighbourhood of each pixel with the center value and considering the result as a binary number (Fig. 6). Then the histogram of the labels can be used as a texture descriptor. Binary codes are so called 'micro-textons' that represent texture primitives such as curved edges, flat or convex areas.

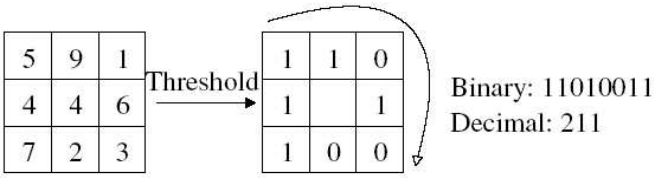

Figure 6: The basic LBP operator [10]

Basic version of LBP uses 3×3 sliding window to code the texture. Recently, the operator has been extended to different sizes and shapes (circular neighbourhood). The size of the neighbourhood directly influences the range of code values. Having operator of size P and radius R, the range of possible codes are from 0 to 2^P. The image texture is described by a 2^P bin histogram of corresponding LBP image.

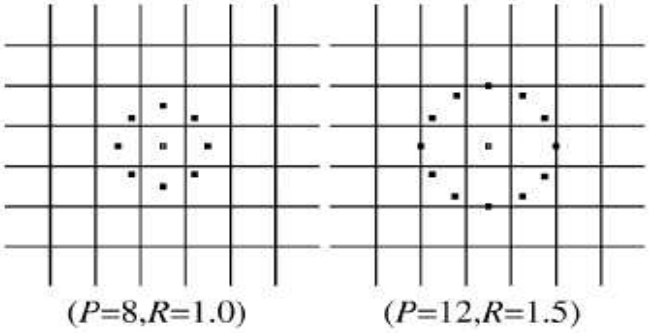

Figure 7: Two examples of the extended LBP [2]: a circular (8; 1) neighbourhood, and a circular (12; 1:5) neighborhood.

Encoding facial texture features can be done in holistic or analytic way. Holistic approach encodes whole face region with LBP features. The disadvantage of this approach is that spatial information about texture is lost. In the second method face region is divided into a grid of patches and each patch is transformed to LBP histogram separately. The latter method encodes the spatial information about the face texture. However, many patches consist of data that is not affected by expression like hair or neck parts. Thus, in this system the LBP operator is applied on two regions that are highly involved in face activity. Those regions are forehead-eyes area and chin-mouth-cheeks area (Fig. 8).

Figure 8: Regions involved in Expression analysis

Before features can be extracted, particular face region need to be normalized. All regions are rescaled to the same size, namely: 90x48 for upper region and 72x48 for lower region. Next, regions are divided into grids of sizes: 4x4 in lower part and 5x4 in upper part of the face. (Fig. 9).

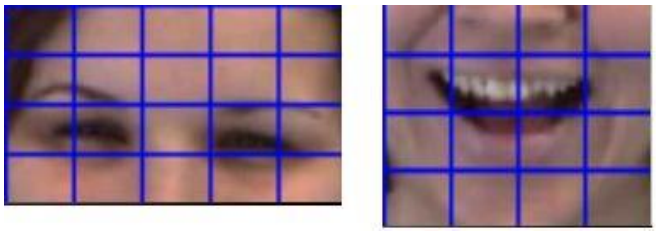

Figure 9: Face region grids

Each window from a grid is encoded with LBP histogram. Basic version of LBP operator was implemented in the proposed system so the amount of bins in the histogram is 2^8=256. Feature vector that represents particular emotion consists of 36 histograms. Thus, each expression is described by 9216 features.

4.3 Expression recognition

The next step in the system is to recognize facial expression based on the extracted features. This task requires classifier training with a set of images with particular emotions displayed.Three principal issues in classification task are: choosing good feature set, efficient machine learning technique and diverse database for training. Feature set should be composed of features that are discriminative and characteristic for particular expression. Machine learning technique is chosen usually by the sort of a feature set. Finally, database used as a training set should be big enough and contain various data.

4.3.1 Databases

One of the most important aspects of developing any new recognition or detection system is the choice of the database that will be used for testing the new system. For the purpose of training, two facial expression databases are to be used. First one is the FG-NET Facial Expression and Emotion Database [4] which consists of MPEG video files with spontaneous emotions recorded. Database contains examples gathered from 18 subjects (9 female

and 9 male). The system has to be trained with captured video frames in which the displayed emotion is very representative. The training set consists of images of seven states neutral and emotional (surprise, fear, disgust, sadness, happiness and anger).

Second database is the Cohn-Kanade Facial Expression Database [5] and contains 486 image sequences displayed by 97 posers. The sequence displays the emotion from the start to the peak, however, only the last image of a sequence is used for training. The training and testing sets formed from Cohn-Kanade database contain images divided into seven classes: neutral, surprise, fear, disgust, sadness, happiness and anger. Having appropriate training sets, the procedure of classifier training can be performed.

4.3.2 Classifier

Many classifiers have been applied to expression recognition such as neural network (NN), support vector machines (SVM), linear discriminant analysis (LDA), K-nearest neighbour, multinomial logistic ridge regression (MLR), hidden Markov models (HMM), tree augmented naive Bayes, and others [11].

In proposed method the Support Vector Machine [12, 13] with Radial Based Kernel Function is used as a classifier. The Support Vector Machine is an adaptive learning system which receives labelled training data and transforms it into higher dimensional feature space. Then separating hyper-plane with respect to margin maximization is computed to determine the best separation between classes. The greatest advantage of SVM is that even with small set of training data it has good performance in generalization. Shan et al. [3] evaluated the performance of LBP features as facial expression descriptors and they obtained the result of 89% by use of SVM with RBF Kernel trained with CK database. Recently Bartlett et al [9] conducted similar experiments. They selected 313 image sequences from the Cohn-Kanade database. The sequences came from 90 subjects, with 1 to 6 emotions per subject. The facial images were converted into a Gabor magnitude representation using a bank of 40 Gabor filters. Then they performed 10-fold cross-validation experiments using SVM with linear, polynomial, and RBF kernels. Linear and RBF kernels performed best, with recognition rates of 84.8% and 86.9% respectively.

As a powerful machine learning technique for data classification, SVM [14] performs an implicit mapping of data into a higher (maybe infinite) dimensional feature space, and then finds a linear separating hyper-plane with the maximal margin to separate data in this higher dimensional space. Given a training set of labelled examples $\{(x_i,y_i),i=1,...,l\}$ where $x_i \epsilon R^n$ and $y_i \epsilon\{1,-1\}$,a new test example x is classified by the following function:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l}(a_i y_i \, \text{K}(x_i, \text{x}) + \text{b}\right)$$

where a_i are Lagrange multipliers of a dual optimization problem that describe the separating hyper-plane, K(.,.) is a kernel function, and b is the threshold parameter of the hyper-plane. The training sample x_i with $a_i > 0$ is called support vectors, and SVM finds the hyper-plane that maximizes the distance between the support vectors and the hyper-plane. SVM allows domain-specific selection of the kernel function. Though new kernels are being proposed, the most frequently used kernel functions are the linear, polynomial, and Radial Basis Function (RBF) kernels.

4.4 Music

The last and the most important part of this system is the playing of music according to the user's current emotional state. Once the face expression of the user has been classified, user's corresponding emotion state is found. A musical database consisting of songs from a variety of domains and pertaining to a number of emotions is maintained .Every song in each emotion category of the database is assigned weights. Also, each emotion detected is assigned weights. The algorithm finds out a song closest in weight with the classified emotion. This song is then played.

5. CONCLUSIONS AND FUTURE WORK

The aim of this paper was to explore the area of automatic facial expression recognition for implementation of an emotion based music player. Beginning with the psychological motivation for facial behavior analysis, this field of science has been extensively studied in terms of application and automation. Manual face analysis used by psychologists was quickly replaced by suitable computer software. A wide variety of image processing techniques was developed to meet the facial expression recognition system requirements. Apart from theoretical background, this work provides ways to design and implement Emotion based music player. Proposed system will be able to process the video of facial behavior, recognize displayed actions in terms of basic emotions and then play music based on these emotions. Major strengths of the system are full automation as well as user and environment independence. Even though the system cannot handle occlusions and significant head rotations, the head shifts are allowed. In the future work, we would like to focus on improving the recognition rate of our system. Also, we would like to develop a mood-enhancing music player in the future which starts with the user's current emotion(which may be sad) and then plays music of positive emotions thereby eventually giving a joyful feeling to the user. Finally, we would like to improve the time efficiency of our system in order to make it appropriate to use in different applications.

Acknowledgement

We would like to express our gratitude to our project guide Prof. Prakash H. Paranjape and also our Head Of Department Prof. N. M. Shekokar for their valuable advice and support. We would also like to thank Prof. T. Kanade, Prof. J.F. Cohn and Prof. Y. Tian for providing the Cohn-Kanade database and Dr.-Ing. F. Wallhoff for

the access to FG-NET Facial Expressions and Emotion Database.

REFERENCES

[1] P. Viola and M. J. Jones, "Robust real-time object detection", International Journal of Computer Vision, Vol. 57, No. 2, p.137–154, 2004.
[2] T. Ojala, M. Pietikainen, T. Menp, "Multiresolution gray-scale and rotation invariant texture with local binary patterns", IEEE transactions on Pattern Analysis and Machine Intelligence Vol. 7, No. 7, p. 971-987, 2002.
[3] C.Shan, S. Gong, P. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study", Image and Vision Computing, Vol. 27, p. 803-816, 2009.
[4] F. Wallhoff, "Facial Expressions and Emotion Database", Technische Universität München, 2006,http://www.mmk.ei.tum.de/~waf/fgnet/feedum.html.
[5] T. Kanade, J. F. Cohn, Y. Tian, "Comprehensive database for facial expression analysis", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, p.46-53, 2000.
[6] Y. Tian, T. Kanade, and J.F. Cohn, Facial Expression Analysis, Handbook of Face Recognition, Springer, October 2003.
[7] Z. Zhang, M. J. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron, in IEEE FG, April 1998.
[8] Y. Tian, "Evaluation of face resolution for expression analysis, in IEEE Workshop on Face Processing in Video, 2004.
[9] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan, "Real time face detection and facial expression recognition: Development and application to human computer interaction, in CVPR Workshop on CVPR for HCI, 2003.
[10] T. Ahonen, A. Hadid, and M. Pietikinen, "Face recognition with local binary patterns," in ECCV, 2004, pp. 469–481.
[11] I. Cohen, N. Sebe, Garg A., L. Chen, and T. Huang, Facial expression recognition from video sequences: Temporal and static modeling," CVIU, vol. 91, pp. 160–187, 2003.
[12] A. Geetha, V. Ramalingam, S. Palanivel and B. Palani- appan, "Facial Expression Recognition—A Real Time Approach," International Journal of Expert Systems with Applications, Vol. 36, No. 1,2009.
[13] J. C. B. Christopher, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery,Vol. 2, No. 2, 1998, pp. 121-167.
[14] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[15] M. Pantic, L. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, 2000.
[16] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression," Proc. Int'l Conf. Automatic Face and Gesture Recognition, p. 396-401, 1998.