Emotion Index of Cover Song Music Video Clips Based on Facial Expression Recognition

Georgios Kavalakis[1], Nikolaos Vidakis[1], and Georgios Triantafyllidis[2(✉)]

[1] Informatics and Multimedia Department, TEI of Crete, Heraklion, Greece
gkavalakis@gmail.com, nv@ie.teicrete.gr
[2] Medialogy Section, ADMT, Aalborg University Copenhagen, Copenhagen, Denmark
gt@create.aau.dk

Abstract. This paper presents a scheme of creating an emotion index of cover song music video clips by recognizing and classifying facial expressions of the artist in the video. More specifically, it fuses effective and robust algorithms which are employed for expression recognition, along with the use of a neural network system using the features extracted by the SIFT algorithm. Also we support the need of this fusion of different expression recognition algorithms, because of the way that emotions are linked to facial expressions in music video clips.

Keywords: Facial expression recognition · SIFT · Cover song video clips

1 Introduction

Human expression recognition is a fundamental problem in computer vision attracting great interest from the research community over the last years. Human expressions involve facial expressions, gestures, voices, etc. In this work, we will focus on the facial expression recognition which is the task of automatically identifying and classifying expressions in an image or video sequence. This is still a difficult task for computer vision to perform, although humans recognize facial expressions without effort or delay.

It is also known that music is a basic way of expressing human emotions. But there is somehow a rather different way of face expressing emotions in music performances and in music video clips compared to the same emotions in everyday life. When someone is singing or playing music, he/she is usually more expressive and also may employ different or more intense or even new expressions for specific emotions. For example, a singer looking down usually indicates a sad expression, while this not always true in everyday life expressions. In this context, this paper considers a specific and interesting case of facial expressions recognition: recognition in music video clips of cover songs, aiming at producing an emotion index, labelling the video clip.

The proposed scheme employs already known and used methods of classifying emotions such as the Logistic Regression (LogReg) [1], the Classification and Regression Trees (CR-tree) [2], Linear discriminant analysis (LDA) [3], k-Nearest

A.L. Brooks and E. Brooks (Eds.): ArtsIT/DLI 2016, LNICST 196, pp. 248–255, 2017.
DOI: 10.1007/978-3-319-55834-9_29

neighbour (k-NN) [4] and Quadratic discriminant analysis (QDA) [5]. The paper also suggests the use of SIFT feature extraction algorithm [6] for emotion classification. SIFT is an algorithm in computer vision that detects and describes local features in images. In this context, for any face in an image, interesting points can be extracted to provide a "feature description" of the face. This description can then be used to recognize the facial emotions in the image with a use of a neural network.

Before processing the facial expressions recognition, an algorithm for face detection should be employed, in order to detect faces that will be evaluated accordingly. However, detecting faces in music video clips is also a challenging task. Face position, lighting, occlusions, video quality are factors that affect the face detection performance.

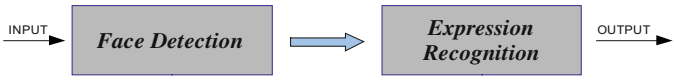


Fig. 1. General scheme

For the facial expression recognition task, we employ a novel scheme with a fusion decision system for each detected face in the music performance. More specifically, it fuses the aforementioned expression recognition techniques for getting a more reliable final decision regarding the emotion index. This fusion of such techniques improves the efficiency of the system, since the facial expressions during performing music may be different or even more difficult to be classified compared to the everyday expressions, because of the way that emotions are linked to facial expressions while performing music.

After this short introduction the rest of the paper is organized as follows: In the following section, we present the methodology and the algorithms of the proposed system. Then, the experimental results for the case of cover song music video clips are presented and finally the conclusions and future work are drawn.

2 Methodology

The program's input is a music video clip of a cover song. The general scheme is illustrated in Fig. 1. A group of frames is selected (e.g. 5 frames per second) and processed with the face detection algorithm. If the detected face is considered as acceptable (see next section for more details) for the expression analysis, this face image is stored for further processing. This includes the expression analysis which fuses different methods and produces a decision for that face image. This decision is an emotion index. If the detected face is not acceptable or there is not any detected face, the algorithm checks the next group of frames until it finds an acceptable face image for processing.

2.1 Face Detection

The first step of the proposed scheme is the detection of faces within the video frames of a musical video clips. A critical issue is the question if the detected faces are suitable for further (expression) analysis. So it is essential that the basic features on the faces (i.e. two eyes and mouth) should be detected, helping in extracting information. In this context, the face detection algorithm of [7] was employed, which uses a cascaded classifier. In this technique of face detection, we also added the constraint of detecting the two eyes and the mouth, since these are the main features which export the information for the facial expression. Once we detect the two eyes and the mouth, this face image is accepted and we proceed to the next step of the expression recognition.

2.2 Expression Recognition

The facial expression recognition has been a subject of research in the computer science for a long time and there is a lot of research on this area. However, the classification rules are somehow different in recognizing expressions of an artist in a music video clip, since an artist playing music or singing a song, presents specific facial expressions, according to his/her emotions, which may differ in a way from the everyday expressions. This fact makes the expression recognition even harder and proves the correctness of the fusion approach which is suggested in this paper. Some more details about the algorithms we used for expression recognition:

- **Logistic Regression (LogReg)** [1]: The logistic regression analysis offers an elegant possibility of examining the influence of several (of quantitative or qualitative) arguments or "factors of risk". The idea of the logistic regression is based on the conception that the probability of an event with an involution model can be functionally described. The influence of the arguments can be modelled directly. A further advantage is that these variables can be usually transferred in their original form to the model. Besides, only the involution coefficients must become estimated, which reduces the number of necessary statistic tests.
- **Classification and Regression Trees (CR-Tree or CART)** [2]: The CR-Tree decision tree is a binary recursive partitioning procedure capable of processing continuous and nominal attributes as targets and predictors. Data are handled in their raw form; no binning is required or recommended. Beginning in the root node, the data are split into two children, and each of the children is in turn split into grandchildren. Trees are grown to a maximal size without the use of a stopping rule; essentially the tree-growing process stops when no further splits are possible due to lack of data. The maximal-sized tree is then pruned back to the root (essentially split by split) via the novel method of cost-complexity pruning. The next split to be pruned is the one contributing least to the overall performance of the tree on training data (and more than one split may be removed at a time). The CR-Tree mechanism is intended to produce not one tree, but a sequence of nested pruned trees, each of which is a candidate to be the optimal tree.

- **Linear discriminant analysis (LDA)** [3]: Linear Discriminant Analysis (LDA) is a method of finding a linear combination of variables which best separates two or more classes. In itself LDA is not a classification algorithm, although it makes use of class labels. However, the LDA result is mostly used as part of a linear classifier. The other alternative use is making a dimension reduction before using nonlinear classification algorithms.
- **k-Nearest neighbour (k-NN)** [4]: is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbour.
- **Quadratic discriminant analysis (QDA)** [5]: is one of the most commonly used nonlinear techniques for pattern classification. In the QDA framework, the class conditional distribution is assumed to be Gaussian, however, with an allowance for different covariance matrices. In such cases, a more complex quadratic boundary can be formed. It is therefore reasonable to believe that QDA better fits the real data structure. However, due to the fact that more free parameters are to be estimated (C covariance matrices, where C denotes the number of classes) compared to those in an LDA-based solution (1 covariance matrix), QDA is more susceptible to the so-called small sample size (SSS) problem where the number of training samples is smaller or comparable to the dimensionality of the sample space.
- **Scale-invariant feature transform (SIFT)** [6]: is an algorithm in computer vision to detect and describe local features in images. The algorithm was published by David Lowe in 1999 and computes scale-space extrema of the space Laplacian, and then samples for each one of these extrema a square image patch. SIFT method is actually proposing descriptors that are invariant to image translations and rotations, to scale changes (blur), and robust to illumination changes. It is also surprisingly robust to large enough orientation changes of the viewpoint (up to 60 degrees). The initial goal of the SIFT method is to compare two images (or two image parts) that can be deduced from each other (or from a common one) by a rotation, a translation, and a zoom [8]. The method turned out to be also robust to large enough changes in view point angle, which explains its success also in object recognition [9]. So, the proposed scheme employs the SIFT method for expression recognition since the SIFT feature keypoints are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images. Therefore, we actually build a neural network and train it by using as input the SIFT feature keypoints of several face images and as output the respective emotion. Regarding this SIFT-based neural network, there is a problem that there are many feature keypoints which are not associated to the expressions, but only with the face characteristics. In this context, the efficient choice of the face regions

is critical. To solve this problem, we may perform a SIFT-based neural network only to the eyes region (SIFT top) and to the mouth region (SIFT down), since these regions are greatly affected by facial expressions (Fig. 2).

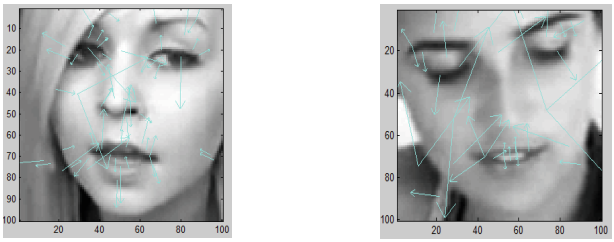


Fig. 2. SIFT features keypoints

The general facial expression flowchart is depicted in Fig. 3. Each one of the faces which were detected in the face detection part of the proposed scheme is analysed by using all six algorithms presented above. These outcomes are fused to result the final decision for the expression recognition of the specific face image which was analysed. Next, we apply the same algorithm to other detected faces and we finally produce the emotion index of the whole video. For simplicity, we have chosen to use the same weight in all six methods.

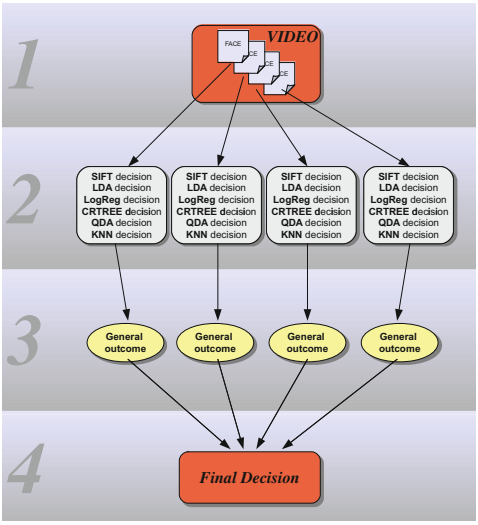


Fig. 3. Emotion recognition analysis

3 Experimental Results

Figure 4 depicts a screenshot of the interface (which is still in a beta version). The user may select the music video clip that will be processed for the facial expression recognition. The interface has been designed in such a way where there is a window playing the video clip, while the faces detected to be suitable for the expression analysis are shown below this video window. The right panel of the interface presents the results (in our proof of concept experiment we have chosen to use on the happy and sad emotions) obtained from each face image according to each algorithm.

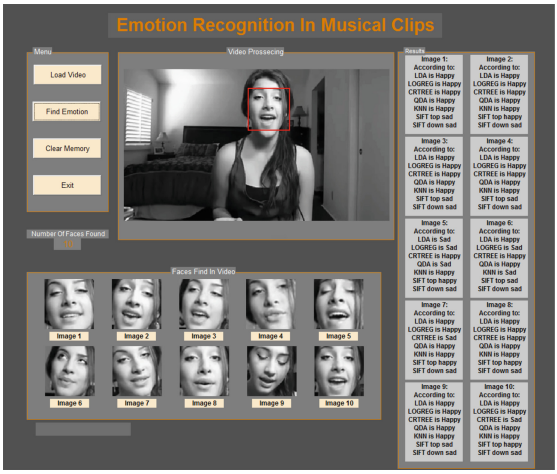


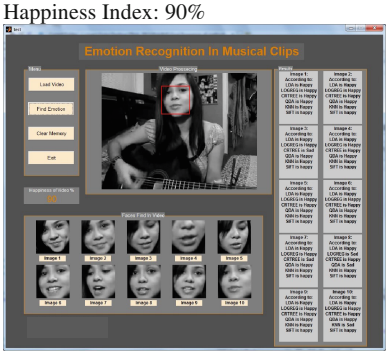
Fig. 4. Program's Interface

So taking into consideration the output of each algorithm, we can reach to a final emotion expression decision for each face image (see Fig. 3). Then, considering these decisions, we can draw a degree (index) of the emotion (e.g. happiness or sadness) for each video (based on the average values and rounded in tens).

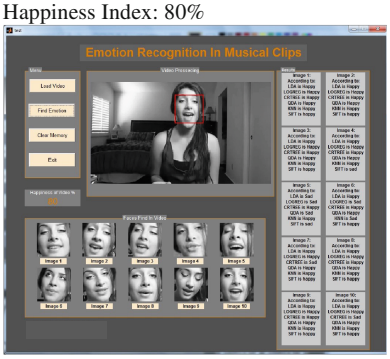
In our experiments we used 4 randomly selected cover song video clips from YouTube. We applied the suggested scheme and concluded to an overall index regarding the emotion of happiness. Results are shown in Table 1. Results greater than 70% indicate a happy song, lower than 30% indicate a sad song, while results around 50% indicate a rather neutral song.

Table 1. Emotion (happiness) index on four YouTube cover song videos

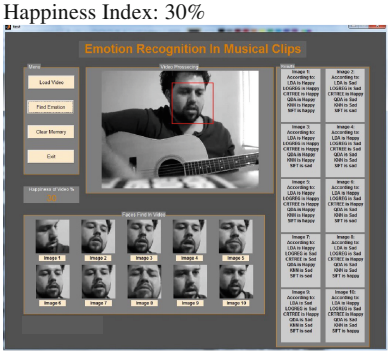
Cover Song 1: Song Title: I am yours
www.youtube.com/watch?v=dQz0U6LV-ME
Happiness Index: 90%



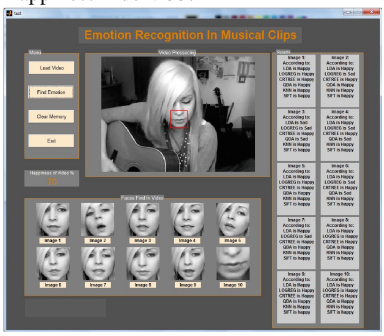
Cover Song 2: Song Title: Just the way you are
www.youtube.com/watch?v=dW6VGm3I8jY
Happiness Index: 80%



Cover Song 3: Song Title: Creep
www.youtube.com/watch?v=Ph1GaDwdecI
Happiness Index: 30%



Cover Song 4: Song Title: The writer
www.youtube.com/watch?v=92dxTXL46h4
Happiness Index: 60%



4 Conclusion and Future Work

Automatic emotion labelling of music video clips based on the facial expression recognition is a challenging task, but also a very useful tool. There are internet services offering huge collections of music video clips that may use such methods for classifying the videos in a data base accordingly. Also users may use the proposed scheme for organizing and sorting their personal collections of music video clips. In this way and depending on the user's mood a music player may automatically choose the right music video clip for playing.

In this context, this paper presents a preliminary study of creating an emotion index of cover song video clips by recognizing facial expressions. Improvements, such as more effective segmentation of areas such as the mouth and the eyes and the selection of proper feature keypoints may produce more accurate SIFT-based neural network's

decisions. Also, the creation of a more effective data base is one of our priorities, aiming to facial expression recognition methods to produce better results. Another direction of future work is the weight-based fusion of the results. Finally, the range of the emotions that can be recognized from the system is about to grow and cover a more realistic range of emotions of a music video clip.

References

1. Agresti, A.: Categorical DATA Analysis. Wiley, New York (1990)
2. Steinberg, D., Colla, P.: CART™ Interface and Documentation. Salford Systems, San Diego (1997)
3. Balakrishnama, S., Ganapathiraju, A.: Linear Discriminant Analysis - A Brief Tutorial. Mississippi State Univ., Starkville (1998)
4. Hall, P., Park, B.U., Samworth, R.J.: Choice of neighbor order in nearest-neighbor classification. Annal. Stat. **36**, 2135–2152 (2008)
5. Wang, J., Plataniotis, K.N., Lu, J., Venetsanopoulos, A.N.: Kernel quadratic discriminant analysis for small sample size problem. Pattern Recogn. **41**(5), 1528–1538 (2008)
6. Lowe D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
7. Kienzle, W., Bakir, G., Franz, M., Scholkopf, B.: Face detection - efficient and rank deficient. In: Advances in Neural Information Processing Systems 17, pp. 673–680 (2005)
8. Foo, J.J., Sinha, R.: Pruning SIFT for scalable near-duplicate image matching. In: Proceedings of the Eighteenth Conference on Australasian Database, vol. 63, pp 63–71 (2007)
9. Moreels P., Perona P.: Common-frame model for object recognition. In: Advances in Neural Information Processing Systems, pp. 953–960 (2004)