Facial Expressions Recognition from Image Sequences

Zahid Riaz, Christoph Mayer, Michael Beetz and Bernd Radig

Department of Informatics, Technische Universität München,
D-85748 Garching, Germany

Abstract. Human machine interaction is one of the emerging fields for the coming years. Interacting with others in our daily life is a face to face interaction. Faces are the natural way of interaction between humans and hence also useful in human machine interaction.
This paper describes a novel technique to recognize the human facial expressions and manipulating this task for human machine interaction. We use 2D model based approach for human facial expression recognition. An active shape model (ASM) is fitted to the face image and texture information is extraced. This shape and texture information is combined with optical flow based temporal information of the image sequences to form a feature vector for the image. We experimented on image sequences of 97 different persons of Cohn-Kanade-Facial Expression Database. A classification rate of 92.4% is obtained using a binary decision tree classifier, whereas a classification rate of 96.4% is obtained using pairwise classifier based on support vector machines. This system is capable to work in realtime.

Key words: Face Modeling, Active Appearance Models, Facial Expressions Recognition, Face Recognition

1 Introduction

Human face analysis plays an important role for the development of social robots. Such scenarios arise when humans and robots are working on some common task. Existing methods for human-machine interaction are often considered unintuitive. Especially for social robots which are operating in various systems in daily life. They require timely human intervention and sometimes result in unfriendly mannered. On the other hand it requires a lot of time and effort to train human to interact with machines which is still not intuitive. However, in this paper we have tried to use those characteristics of the humans which could be intuitively interpreted by the machines and hence resulting in a an autonomous system working without human interaction. To participate in natural human-machine interaction, machines must be able to derive information from human communication channels, such as spoken language, gestures or facial expressions. We evaluate human face information in order to classify the facial expressions which could be easily embedded in advanced systems to study human facial behaviors. We used model based image interpretation techniques to extract information about facial changes. Models impose knowledge about the object of interest and reduce high dimensional image data to a small number of expressive model parameters. The model used by our system is a point distribution model (PDM) [**?**]. A combination of the facial features is used for binary decision tree to classify six basic facial expressoins i.e.

anger, fear, surprise, saddness, laugh and disgust. Unlike many competing systems, the approach is fully automated and requires no human intervention.

This work focuses on one of the aspects of natural human-computer interfaces: our goal is to build a real-time system for facial expressions recognition that could robustly run in real-world environments. We develop it using model-based image interpretation techniques, which have proven its great potential to fulfill current and future requests on real-world image understanding.

2 Related Work

We explain facial expressions recognition as a three step process consisting of face detection, features extraction, and expressions classification [**?**]. The first step aims at determining the position and shape of the face in the image by fitting a model. Features descriptive for facial expressions or head gestures are extracted in the second step. In the third step a classifier is applied to the features to identify the expressions.

Cootes et al. [**?**] introduced modelling shapes with Active Contours. Further enhancements included the idea of expanding shape models with texture information [**?**]. In order to fit a model to an image, Van Ginneken et al. learned local objective functions from annotated training images [**?**]. In this work, image features are obtained by approximating the pixel values in a region around a pixel of interest The learning algorithm used to map images features to objective values is a k-Nearest-Neighbor classifier (kNN) learned from the data. We used similar methodology developed by Wimmer et al. [**?**] which combines multitude of qualitatively different features [**?**], determines the most relevant features using machine learning [**?**], and learns objective functions from annotated images [**?**].

In order to find features, Michel et al. [**?**] extracted the location of 22 feature points within the face and determine their motion between an image that shows the neutral state of the face and an image that represents a facial expression. The very similar approach of Cohn et al. [**?**] uses hierarchical optical flow in order to determine the motion of 30 feature points.

A set of training data hence formed from the extracted features is learned on a classifier. Some approaches infer the facial expression from rules stated by Ekman and Friesen [**?**]. This approach is applied by Kotsia et al. [**?**] to design Support Vector Machines (SVM) for classification. Michel et al. [**?**] train a Support Vector Machine (SVM) that determines the visible facial expression within the video sequences of the Cohn-Kanade Facial Expression Database by comparing the first frame with the neutral expression to the last frame with the peak expression. The features we used here are quite efficient not only for facial expression recognition but also for face recognition against facial expressions [**?**].

3 Our Approach

This paper describes the classification of facial expresions for the image sequences. We derive features which are suitable not only for person identity but also for classifying six basic facial expressions. The face feature vector consists of the shape, texture and

temporal variations, which sufficiently defines global and local variations of the face. All the subjects in the database are labeled for different facial expressions. A fitted face model, on the training images is then used for defining the reference shape in our experiments. This reference shape is calculated by finding the mean shape of all the shapes in the database.

In this paper a robustly fitted model is coupled with temporal information of the image sequences. A point distribution model (PDM) consisting of 134 landmarks defined on various location on the face is used as an active shape model. An objective function is learned for fitting this model to the faces. After fitting the model to the example face image, texture information is extracted from the example image on a reference shape which is the mean shape of all the shapes of database. Image texture is extracted using planar subdivisions of the reference and the example shapes. Texture warping between the subdivisions is performed using affine transformation. Principal Components Analysis (PCA) is used to obtain the texture and shape model parameters of the example image. This approach is similar to extracting Active Appearance Model (AAM) parameters. In addition to AAM parameters, temporal features of the facial changes are also calculated. Local motion of the feature points is observed using optical flow. We use reduced descriptors by trading off between accuracy and runtime performance. These features are then used for binary decision tree (BDT) for facial expressions recognition. A detailed process flow of our approach is shown in Figure **??**. Our approach achieves real-time performance. This computer vision task comprises of various phases for which it exploits model-based techniques that accurately localize facial features, seamlessly track them through image sequences, and finally infer facial expressions.

The remainder of this paper is divided in four sections. In section 4 model based image interpretation is described along with the model fitting process. Sections 5 discusses about the model based feature extraction technique comprising of shape and appearance along with the temporal features. Section 6 deals with feature classification along with experimentations and finally the conclusions in section 7.

4 Model-based Image Interpretation

Model-based techniques consist of four components: the model, the initialization algorithm, the objective function, and the fitting algorithm. Our approach makes use of a statistics-based deformable model, as introduced by Cootes et al. [**?**]. The model contains a parameter vector $\boldsymbol{p}$that represents its possible configurations, such as position, orientation, scaling, deformation and texture. Models are mapped onto the surface of an image via a set of feature points, a contour, a textured region, etc. Referring to [**?**], deformable models are highly suitable for analyzing human faces with all their individual variations. Its parameters p comprise the translation, scaling factor, rotation, and a vector of deformation parameters $\boldsymbol{b}$. The latter component describes the configuration of the face, such as the opening of the mouth, roundness of the eyes, raising of the eye brows.

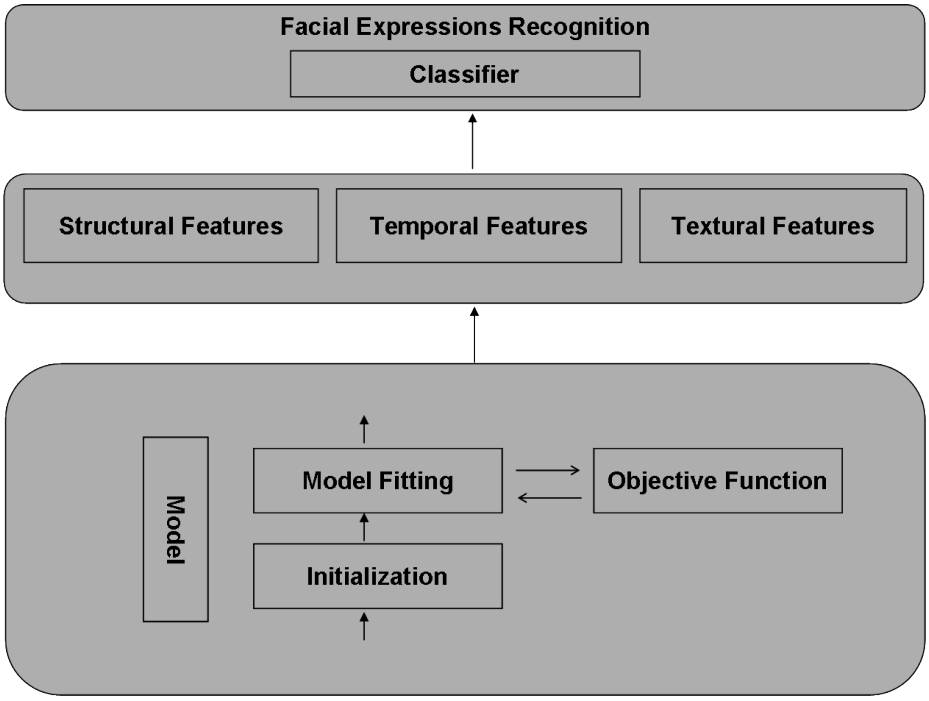

Fig. 1. Our Approach: Different modules working independently towards facial expressions recognition.

4.1 Model Initialization

The initialization algorithm automatically localizes the object roughly to interpret and computes an initial estimate of the translation and scaling parameters. We integrate the approach of Viola and Jones to detect frontal faces. In order to obtain higher accuracy, we again apply this algorithm to detect facial components, such as eyes and mouth within the face bounding box and derive a rough estimation of the shape parameters as well. Usually resulting detectors are not able to robustly localize the eyes or the mouth in a complex image, because it usually contains a lot of information that was not part of the training data. However, it is highly appropriate to determine the location of the eyes within a pure face image or within the face region of a complex image.

4.2 Model fitting

The model parameters determined in the initialization step are only a rough guess and therefore need refinement. An objective function f(I, p) that yields a comparable value to measure how accurately a parameterized model p describes an image I is utilized in this challenge. It is also known as the likelihood, similarity, energy, cost, goodness, or quality function. Without losing generality, we consider lower values to denote a better model fit.

Traditionally, objective functions are manually specified by first selecting a small number of simple image features, such as edges or corners, and then formulating mathematical calculation rules. Afterwards, the appropriateness is subjectively determined by inspecting the result on example images and example model parameterization. If the result is not satisfactory the function is tuned or redesigned from scratch. This heuristic approach relies on the designer's intuition about a good measure of fitness. Earlier works by Wimmer et al. [**?**] show that this methodology is erroneous and tedious.

The fitting algorithm searches for the model that best describes the face visible in the image. Therefore, it aims at finding the model parameters that minimize the objective function. Fitting algorithms have been the subject of intensive research and evaluation, e.g. Simulated Annealing, Genetic Algorithms, Particle Filtering, RANSAC, CONDENSATION, and CCD, see [16] for a recent overview and categorization. We propose to adapt the objective function rather than the fitting algorithm to the specifics of our application. Therefore, we are able to use any of these standard fitting algorithms, the characteristics of which are well-known, such as termination criteria, runtime, and accuracy.

To avoid these drawbacks, we recently proposed an approach that learns the objective function from annotated example images [**?**]. It splits up the generation of the objective function into several tasks partly automated. This provides several benefits: firstly, automated steps replace the labor-intensive design of the objective function. Secondly, this approach is less error prone, because giving examples of good fit is much easier than explicitly specifying rules that need to cover all examples. Thirdly, this approach does not rely on expert knowledge and therefore it is generally applicable and not domain-dependent. The bottom line is that this approach yields more robust and accurate objective functions, which greatly facilitate the task of the fitting algorithm.

5 Face Model based Feature Extraction

In order to classify facial expressions, we require a reasonable set of descriptive features. We apply two kinds of facial features to represent the face state and face movement: global features and local features. The first represent face properties of single images, such as the opening of the mouth or rising of the eye-brows. We combine them with information about movement within the face region to form an invariant feature vector. As our experimental evalution will prove, these extracted features are efficient for face and facial expressions recognition.

The shape $\boldsymbol{x}$ is parameterized by using mean shape $\boldsymbol{x_m}$ and matrix of eigenvectors $\boldsymbol{P_s}$ to obtain the parameter vector $\boldsymbol{P_s}$ [**?**].

$$\boldsymbol{x} = \boldsymbol{x_m} + \boldsymbol{P_s b_s} \tag{1}$$

Shape information is extracted to reflect changes in facial expression. This information determines the state of various facial features such as eye brows or the mouth. For texture extraction, we calculate a reference shape which we chose to be the mean shape, obtained by taking the mean of all the shapes of all persons in our database. Figure **??** shows shape information interpretation.

Fig. 2. Shape information is extracted to represent the state of various facial feature such as the rising of the eye brows or the opening angle of the mouth.

Since the points distribution defines a convex hull of points in space so a suitable planar subdivision could be defined for the reference shape to map image texture. Delaunay triangulation is used to divide the shape into a set of distinct facets. The extracted texture is parameterized using PCA by using mean texture $\boldsymbol{g_m}$and matrix of eigenvectors $\boldsymbol{P_g}$to obtain the parameter vector $\boldsymbol{b_g}$ [**?**]. Figure **??** shows texture extracted from face image.

$$\boldsymbol{g} = \boldsymbol{g_m} + \boldsymbol{P_g b_g} \tag{2}$$

Fig. 3. Texture information is represented by an appearance model. Model parameters of the fitted model are extracted to represent single image information.

Since facial expressions emerge from muscle activity, the motion of particular feature points within the face gives evidence about the facial expression. These features further help the classifier to learn the motion activity and their relative location is connected to the structure of the face model. Note that we do not specify these locations manually, because this assumes a good experience of the designer in analyzing facial expressions. In contrast, we automatically generate T feature points that are uniformly distributed. We expect these points to move descriptively and predictably in the case of a particular facial expression. We sum up the motion $t_{x,i}$ and $t_{y,i}$ of each point $1 \leq i \leq T$

during a short time period. We set this period to 2 sec to cover slowly expressed emotions as well. The motion of the feature points is normalized by the affine transformation of the entire face in order to separate the facial motion from the rigid head motion. In order to determine robust descriptors, PCA determines the H most relevant motion patterns (principal components) visible within the set of training sequences. A linear combination of these motion patterns describes each observation approximately correct. This reduces the number of descriptors by enforcing robustness towards outliers as well. As a compromise between accuracy and runtime performance, we set the number of feature points to T = 140 and the number of motion patterns b_t to H = 14 containing. Figure **??** shows motion pattern withing an image.

Fig. 4. Motion patterns within the image are extracted and the temporal features are calculated from them. These features are descriptive for a sequence of images rather than single images.

We compile all extracted features into a single feature vector. Single image information is considered by the structural and textural features whereas image sequence information is considered by the temporal features. The overall feature vector then becomes:

$$\boldsymbol{u} = (\boldsymbol{b_{s1}}, ..., \boldsymbol{b_{sm}}, \boldsymbol{b_{t1}}, ..., \boldsymbol{b_{tm}}, \boldsymbol{b_{g_1}}, ..., \boldsymbol{b_{g_m}}) \tag{3}$$

Where $\boldsymbol{b_s}$, $\boldsymbol{b_t}$ and $\boldsymbol{b_g}$ are shape, temporal and textural parameters respectively.

6 Experimental Evaluation

Experiments have been performed on Cohn-Kanade-Facial-Expression database (CKFE-DB) for human faces. The CKFE-DB contains 488 short image sequences of 97 different persons performing the six universal facial expressions [**?**]. It provides researchers with a large dataset for experimenting and benchmarking purpose. Each sequence shows a neutral face at the beginning and then develops into the peak expression. Therefore, a decision has to be made at which point the face changes from a neutral state to an acctual expression and we label the first half of every sequence to be neutral and the second half to depict expression.

Note that this database does not contain natural facial expressions, but volunteers were asked to act. Furthermore, the image sequences are taken in a laboratory environment with predefined illumination conditions, solid background and frontal face views. Algorithms that perform well with these image sequences are not immediately appropriate for real-world scenes.

We apply machine learning techniques to derive calculation rules for facial expression that are calculated from the complete assembled feature vector as well as from subsets of the features. The classifiers determine the facial expression for every image separately. Note that the information provided by the temporal features is still dependent on the predecent images and therefore takes into account not only the information of one image but of all previous images in the sequence. However, In order to avoid overfitting, we used 10-fold cross validation in both cases.

We apply support vector machines (SVM) for pairwise classification. In a first experiment SVMs is trained with the temporal features only and achieved an accuracy of 79.5%. This result is improved to 92.2% by additionally providing the structural features. This is due to the fact that these features provide information about both, the single image and the image sequence. The best result is gained by providing full set of features and the accuracy raised to 96.4%

In order to create a classifier that is applicable in real-world environments as well, we train a classifier that is both, fast and capable of considering several class labels in the data. We apply decision trees for this task and provide them with the completely assembled data vector. The accuracy measured is 92.6%. Note that in contrast to the SVMs the decision tree does not utilize previous information on the image label. Since its accuracy is only 3.8% below the best accuracy of SVM, this proves the high descriptive strength of the features provided.

7 Conclusions

We introduced an idea to develop a feature vector which consists of three types of facial variations. The features extracted from the images are descriptive for both, single images and image sequences. A model based approach is applied to extract shape information and forms the basis for the extraction of textural information and motion information. The assembled features are provided to train a support vector machine that derives the facial expression currently visible. Our experimental evaluation demonstrates that the classifiers are well able to handle this task and therefore proves that the extracted features are higly descriptive. We demonstrated that on two different classifiers, support vector machines and binary decision trees. However the benchmarked database consists of only frontal view of faces. In the future, experiments will be performed to test the capability of the system to work in real time environment.

References

1. M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

2. Edwards, G. J., Taylor, C. J., Cootes, T. F., Interpreting Face Images using Active Appearance Models In *Proceedings of International Conference on Automatic Face and Gesture Recognition* (1998) 300-305.
3. M. Wimmer, Z. Riaz, C. Mayer and B. Radig, Recognizing Facial Expressions Using Model-Based Image Interpretation. *Advances in Human-Computer Interaction*.
4. Wimmer M, Stulp F, Tschechne S, and Radig B, Learning Robust Objective Functions for Model Fitting in Image Understanding Applications. In *Proceedings of the 17th British Machine Vision Conference* pp1159–1168, BMVA, Edinburgh, UK, 2006.
5. Tim F. Cootes and Chris J. Taylor. Active shape models – smart snakes. In *Proceedings of the 3rd British Machine Vision Conference* pages 266 - 275. Springer Verlag, 1992.
6. Cootes T. F., Edwards G. J., Taylor C. J. Active Appearance Models. In *Proceedings of European Conference on Computer Vision* Vol. 2, pp. 484-498, Springer, 1998.
7. Stan Z. Li and A. K. Jain. Handbook of Face recognition *Springer* 2005
8. Kanade, T., Cohn, J. F., Tian, Y. (2000). Comprehensive database for facial expression analysis In *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FGR00)*, Grenoble, France, 46-53.
9. P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, 1978.
10. P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *Fifth International Conference on Multimodal Interfaces*, pages 258–264, Vancouver, 2003.
11. J. Cohn, A. Zlochower, J.-J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396 – 401, April 1998.
12. I. Kotsia and I. Pitaa. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transaction On Image Processing*, 16(1), 2007.
13. Riaz Z. et al. A Model Based Approach for Expression Invariant Face Recognition In *3rd International Conference on Biometrics*, IAPR, Italy, June 2009
14. B. Ginneken, A. Frangi, J. Staal, B. Haar, and R. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.
15. S. Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy*. PhD thesis, University of Basel, Computer Science Department, Basel, CH, January 2005.