Multi-Feature Fusion in Advanced Robotics Applications

Zahid Riaz, Christoph Mayer, Michael Beetz, Bernd Radig

Institut für Informatik
Technische Universität München
D-85748 Garching, Germany
{riaz,mayerc,beetz,radig}@in.tum.de

Saquib Sarfraz

Computer Vision Research Group (COMVis)
Department of Electrical Engineering, COMSATS Institute of Information Technology, M.A.Jinnah Campus, Defense Road off Raiwind Road, Lahore, Pakistan
ssarfraz@ciitlahore.edu.pk

ABSTRACT

This paper describes a feature extraction technique from human face image sequences using model based approach. We study two different models with our proposed approach towards multi-feature extraction. These features are efficiently used for human face information extraction for different applications. The approach follows in fitting a model to face image using robust objective function and extracting textural and temporal features for three major applications naming 1) face recognition, 2) facial expressions recognition and 3) gender classification. For experimentation and comparative study of our multi-features over two models, we use same set of features with two different classifiers generating promising results to explain that extracted features are strong enough to be used for face image analysis. Features goodness has been investigated on Cohn Kanade Facial Expressions Database (CKFED). The proposed multi-features approach is automatic and real time.

Keywords: Feature extraction, Active shape models, Active appearance models, Wireframe model, Face image analysis, Human robot interaction.

1. INTRODUCTION

Human face analysis has been one of the challenging fields over the last few years, especially in the situations where humans are interacting with autonomous machines. These autonomous machines or assistive robots should be intelligent enough to friendly interact with humans. The intuitive approach is to train these robots so that they are able to behave in a natural manner that humans use with each other in daily life interaction. Analyzing human faces for human robot interaction is a challenging task in computer vision application. Generally, in computer vision human face features are not robust to variations like varying head poses, different illumination levels, aging factors, occlusions and make ups. Many researchers have solved the problems of analyzing the faces in constrained environments [1] whereas some solved the problem by considering these variations [2]. In the recent decade model based image analysis of human faces has become a challenging field due to its capability to deal with the real world scenarios. Further it outperforms the previous techniques which were constrained to user intervention with the system either to manually interact with system or to be frontal to the camera. Currently available model based techniques are trying to deal with some of the future challenges like developing state-of-the-art algorithms, improving efficiency, fully automated system development and versatility under different applications.

Model-based image interpretation techniques extract information about facial expression, person identity and gender classification from images of human faces via facial changes. Models take benefit of the prior knowledge of the object shape and hence try to match themselves with the object in an image for which they are designed. Face models impose knowledge about human faces and reduce high dimensional image data to a small number of expressive model parameters. The model parameters together with extracted texture and motion information are utilized to train classifiers that determine person-specific information. A combination of different facial features is used for classifiers to classify six basic facial expressions i.e. anger, fear, surprise, sadness, laugh and disgust, facial identity and gender classification. We refer these features as multi-features and prove their sufficiency for face image interpretation. A multi-feature vector for each image consists of structural, textural and temporal variations of the faces in the image sequence. Shape and textural parameters define active appearance models (AAM) [3]. Temporal features are extracted using optical flow. These extracted multi-features are more informative than AAM parameters since we consider local motion patterns in the image sequences in the form temporal parameters.

In this paper, we emphasize on an automatic multi-feature extraction technique which is useful for various applications considering facial expression variations on the same time. Multi-features correspond to combination of different features variations which sufficiently describe a face image as a non-rigid deformable object. Further these features behave in a natural way where humans extract face identity, expressions and gender in a very first glance. We focus on feature extraction technique which is fully automatic and versatile enough for different applications like face recognition, facial expressions recognition and gender classification. These capabilities of the system suggest applying it in interactive scenarios like human machine interaction, security of personalized utilities like tokenless devices, facial analysis for person behavior and person security. We apply a model based approach to human face image analysis and extract the features which are successfully used for information extraction from the image sequences. This face model consists of point distribution in

2D and consists of shape and texture information along with temporal information for each image in the database. On the other hand, 3D model is a wireframe model called Candide-III [6] consisting of 113 points. We test these features on Cohn Kanade Facial Expression Database (CKFED) for different applications.

The remainder of this paper is divided in four sections. Section 2 explains the related research work for model based approaches, including 2D and 3D models. Section 3 describes in detail the approach followed in this paper for feature extraction and their combination. Section 4 describes feature extraction technique in our case, which is used in section 5 for experimentation. In the last section, we describe the analysis of our technique along with the future extension of this research work. The approach followed in this paper is same for 2D and 3D model unless explained explicitly.

2. RELATED WORK

In the recent decade advancement in the field of camera technology, their mountability on mobiles and availability of high computational powers in personal computers have increased the demand of the user to extract more information from the images for higher applications. For instance, face detection cameras, smile-capture cameras, face recognition in notebooks and picasa by google for tagging face images. These applications not only apply to the face image analysis in challenging environments but also emphasize on insufficiency of the traditional approaches for face image analysis. For instance, traditional face recognition systems have the abilities to recognize the human using various techniques like feature based recognition, face geometry based recognition, classifier design and model based methods [1] but on the other hand similar features are not sufficient for gender recognition or facial expressions recognition. A proposed solution to this problem is to model human faces with the descriptors which contain maximum information perceived by humans. In this regard, model based approaches have been very successful over last few years. Currently the available models used by the researchers are deformable models, point distribution models, rigid models, morphable models and wireframe models [2]. A well-known 2D model using shape and texture parameters is called active appearance model (AAM), introduced by Edwards and Cootes [3]. This model is further used by researchers in various applications like medical imaging [4], its extensions to 2D+3D [5] and 3D analysis of face images [6]. We use a 2D active shape model (ASM) [7] as a base line model in our case which is fitted to face image and texture information is extracted using affine transform. In order to fit a model to an image, Van Ginneken et al. learn local objective functions from annotated training images [8]. In this work, image features are obtained by approximating the pixel values in a region around a pixel of interest. The learning algorithm uses to map images features to objective value is a k-Nearest-Neighbor classifier (kNN) learned from the data. We use a methodology developed by Wimmer et al. [9] which combines multitude of qualitatively different features [10], determines the most relevant features using machine learning and learns objective functions from annotated images [8]. In order to perform face recognition applications many researchers have applied model based approach. Edwards et al [11] use weighted distance classifier called Mahalanobis distance measure for AAM parameters. However, they isolate the sources of variation by maximizing the inter class variations using Linear Discriminant Analysis (LDA), a holistic approach which was used for Fisherfaces representation [12]. However they do not discuss face recognition under facial expression. Riaz et al [13] apply similar features for explaining face recognition using Bayesian networks. However results are limited to face recognition application only. They used expression invariant technique for face recognition, which is also used in 3D scenarios by Bronstein et al [14] without 3D reconstruction of the faces and using geodesic distance. In order to find features for facial expressions, Michel et al [15] extract 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 [16] uses hierarchical optical flow in order to determine the motion of 30~feature points. Some approaches infer the facial expression from rules stated by Ekman and Friesen [17]. This approach is applied by Kotsia et al [18] to design Support Vector Machines (SVM) for classification. Michel et al [15] also train an 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.

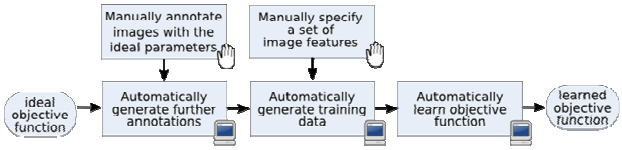


Figure 1. Learning robust objective function [9]

3. OUR APPROACH

In this section we explain our approach including model fitting, image warping and parameters extraction for shape, texture and temporal information. A detailed overview of aour approach is shown in Figure 2.

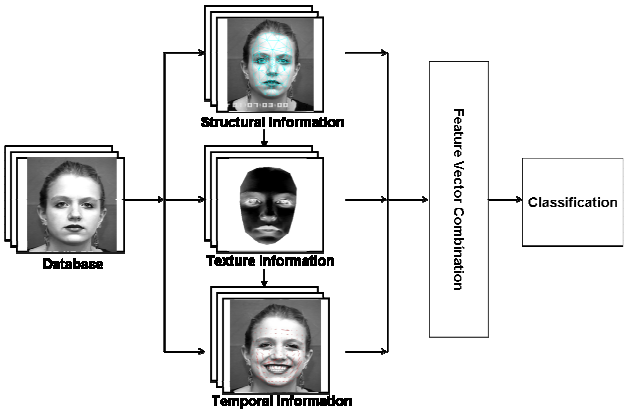


Figure 2. Our approach explaining different modules working towards features extraction.

We use a point distribution model (PDM) for shape which consists of 134 anatomical landmarks in 2D and develop a base line for texture extraction. We use robust objective function for

model fitting. A detailed process flow for learning objective function is shown in Figure 1. In case of 3D, we use wireframe model Candide-III. This model is projected to 2D image using camera transformation and treated as 2D model for texture extraction. The feature vectors extracted from 3D model are two and a half dimensional in nature, since shape information is extracted in 3D space whereas texture is considered 2D. The remaining temporal information is same for both models because we define limited feature points to observe the local motions of the facial features.

After fitting the model to the example face image, texture information is projected from the example image on a reference shape which is the mean shape of all the shapes available in 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 Component Analysis (PCA) is used to obtain the texture and shape parameters of the example image. This approach is similar to extracting 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 run time performance. These features are then used for classifiers for classification. Our approach achieves real-time performance and provides robustness against facial expressions for real-world applications. This computer vision task comprises of various phases shown in Figure 2 for which it exploits model-based techniques that accurately localize facial features, seamlessly track them through image sequences, and finally infer facial features. We specifically adapt state-of-the-art techniques to each of these challenging phases.

4. FEATURE EXTRACTION

Each image in the sequence is used to extract a set of representative features. We start with the model fitting and extract the shape parameters by optimizing local objective function [9]. The shape x is parameterized by using mean shape x_m and matrix of eigenvectors P_s to obtain the parameter vector b_s [19].

$$x = x_m + P_s b_s \quad (1)$$

Figure 3. Model fitted to face images: 3D wireframe model candied-III projected to image plane and fitted to face image (left), 2D point distribution model (PDM) fitted to face image.

Shape information is extracted to reflect changes in facial expressions. 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 is chosen to be the mean shape, obtained by taking mean of all the shapes of all persons in our database. Figure 3 shows shape information interpretation and model fitting results to face image. Shape information is extracted to represent the state of various facial features such as the rising height 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 we use delaunay triangulation to divide the shape into a set of distinct facets. The extracted texture is parameterized using PCA by using mean texture g_m and matrix of eigenvectors P_g to obtain the parameter vector b_g [19]. Figure 4 shows texture extracted from face image.

$$g = g_m + P_g b_g \quad (2)$$

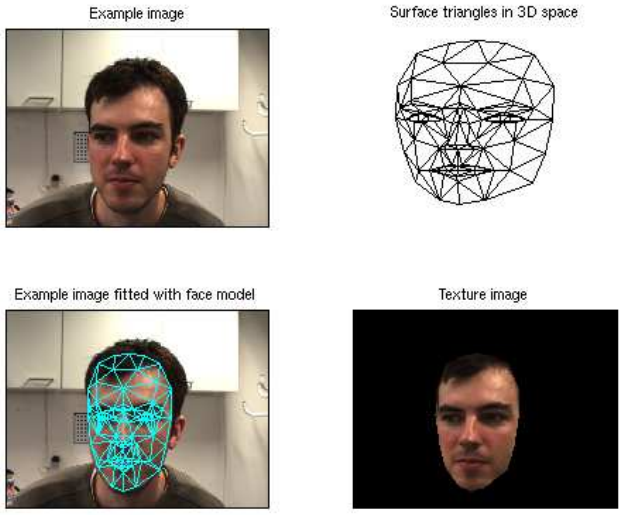


Figure 4. Texture information is represented by an appearance model. Model parameters of the fitted model are extracted to represent single image information.

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 points manually, because this assumes a good experience of the designer in analyzing facial expressions. In contrast, we automatically generate G 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 $g_{x,i}$ and $g_{y,i}$ of each point $1 \leq i \leq G$ 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 run time performance, we set the number of feature points to G = 140 and

the number of motion patterns b_t to H = 14 containing. Figure 5 shows motion pattern within face region.

We combine 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 becomes:

$$u = (b_{sa},\ldots\ldots,b_{sm},b_{ga},\ldots\ldots,b_{gn},b_{ta},\ldots\ldots,b_{tp}) \quad (3)$$

Where b_s, b_t and b_g are shape, temporal and textural parameters respectively.

Figure 5. 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.

5. EXPERIMENTATION

For experimentation purposes, we benchmark our results on Cohn Kanade Facial Expression Database (CKFED). The database contains 488 short image sequences of 97 different persons performing six universal facial expressions [20]. 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. Furthermore, a set of action units (AUs) has been manually specified by licensed Facial Expressions Coding System (FACS) [17] experts for each sequence. 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. We use a subset of this database for our experiments.

For face recognition in the presence of strong facial expressions, we use our extracted features set with Bayesian networks (BN) and binary decision tree (BDT). The face images are frontal however, there exists slight variations in upright position of the head, but the features work well in this case because all the images have been normalized to a standard shape. The face recognition rate is recorded to 98.69% with BN and 93.65% for BDT with 10-fold cross validation. We used standard classifiers from weka [21] for quick experimentation. We repeat the same experiment but apply 3D model features this time and obtain improved results for BDT classifier. Face recognition rate is attained up to 98.49% in case of BDT and 96.42% in case of BN. Figure 6 and Figure 7 show the result for each individual subject in our database in the form of true positives and false positive rates. A slight fall in the case Bayesian network in case of 3D model can be seen from figures and Table 1 and Table 2. The probable reason for this fall is limited texture information in case of 3D model.

In the second experiment for facial expressions recognition, we use extracted features with same classifiers as for face recognition. The extracted features normalize the face image to a standard shape to reduce the facial expressions variations. However temporal and textural parameters contain sufficient facial expressions information. The facial expressions recognition rate is recorded up to 98.30% and 92.63% for boosted BDT and BN respectively in case of 2D, with very less confusion between mostly confused facial expressions like laugh and fear. The results improved in case of 3D model using same classifiers. The CKFE dataset consists majorly of female face images. In order to confirm the feature versatility on gender classification, we have trained the feature set on same classifiers again. The classification rate calculated is 97.39% and 97.88% for BDT and boosted BN classifiers respectively. The results improved in case of 3D and are shown in detail in Table 1 and Table 2.

6. CONCLUSIONS

We introduced a technique to develop a set feature vectors which consist of three types of facial information and its robustness against the expressions changes in human faces. Same features set is applied to three different applications: face recognition, facial expressions recognition and gender classification, which produced the reasonable results in all three cases for CKFED. The results obtained from 3D model are comparatively better to that of 2D model however there is a slight fall in case of face recognition with BN and facial expressions recognition with BDT. The major reason is that the current models have texture information in 2D and 3D model contain only shape as additional information in 3D. Due to restricted frontal face views, it is not possible to extract detailed texture. Face poses are fixed which do not allow capturing full 3D motion of the faces. We consider different classifiers in weka for checking the versatility of our extracted features. The database consists of frontal views with uniform illuminations. This approach can be applied to capture detailed texture and considering head poses and lighting variations. Since the algorithm is working in real time, hence it is suitable to apply it in real time environment keeping in consideration the limitation of database.

7. REFERENCES

[1] W. Zhao, R. Chellapa, A. Rosenfeld and P.J. Philips, Face Recognition: A Literature Survey, UMD CFAR Technical Report CAR-TR-948, 2000.

[2] W. Zhao, R. Chellapa, Face Processing: Advanced Modeling and Methods, In Elsevier Inc. 2006.

[3] Cootes T. F., Edwards G. J., Taylor C. J. Active Appearance Models. European Conference on Computer Vision, Vol. 2, pp. 484-498, Springer, 1998.

[4] Stegmann M. B. Active Appearance Models: Theory Extensions and Cases. In Master Thesis, Technical University of Denmark} 2000.

[5] Xiao, J. Baker, S. Matthews, I. Kanade, T. Real-Time Combined 2D+3D Active Appearance Models, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June, 2004, pp. 535 – 542

[6] Ahlberg J. An Experiment on 3D Face Model Adaptation using the Active Appearance Algorithm. Image Coding Group, Deptt of Electric Engineering Linköping University.

[7] Cootes T. F. and Taylor C. J., Active shape models -- smart snakes. Proceedings of the 3rd British Machine Vision Conference} pages 266 - 275. Springer Verlag, 1992.

[8] Ginneken B., Frangi A, Staal J, Haar B, and Viergever R. Active shape model segmentation with optimal features. IEEE Transactions on Medical Imaging, 21(8):924--933, 2002.

[9] Wimmer M, Stulp F, Tschechne S, and Radig B, Learning Robust Objective Functions for Model Fitting in Image Understanding Applications. Proceedings of the 17th British Machine Vision Conference, pp1159--1168, BMVA, Edinburgh, UK, 2006.

[10] S. Romdhani, Face Image Analysis using a Multiple Feature Fitting Strategy, PhD thesis, University of Basel, Computer Science Department, Basel, CH, January 2005.

[11] Edwards G J, Cootes T F and Taylor C J, Face Recognition using Active Appearance Models, in Proceeding of European Conference on Computer Vision 1998 vol. 2, pp-581-695, Springer 1998.

[12] Belheumeur P N, Hespanha J P and Kreigman D J, Eigenfaces vs Fisherfaces: Recognition using Class Specific Linear Projection, IEEE Transaction on Pattern Analysis and Machine Intelligence} Vol 19, No. 7, July 1997.

[13] Riaz Z. et al. A Model Based Approach for Expression Invariant Face Recognition, 3rd International Conference on Biometrics, Italy, June 2009.

[14] Bronstein A, Bronstein M, Kimmel R, Spira A, 3D face recognition without facial surface reconstruction, In Proceedings of European Conference of Computer Vision, Prague, Czech Republic, May 11-14, 2004

[15] Michel P and Kaliouby R E, Real time facial expression recognition in video using support vector machines. In Fifth International Conference on Multimodal Interfaces, pages 258--264, Vancouver, 2003.

[16] Cohn J, Zlochower A, Lien J, and Kanade T, 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.

[17] Ekman P and Friesen W, The Facial Action Coding System: A Technique for The Measurement of Facial Movement, Consulting Psychologists Press, San Francisco, 1978.

[18] Kotsia I and Pitaa I. Facial expression recognition in image sequences using geometric deformation features and support vector machines, IEEE Transaction On Image Processing, 16(1), 2007.

[19] Li Z. S and Jain A. K, Handbook of Face recognition, Springer 2005.

[20] 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.

[21] Ian H. Witten and Eibe Frank, Data Mining: Practical machine learning tools and techniques Morgan Kaufmann, 2nd Edition, San Francisco, 2005.

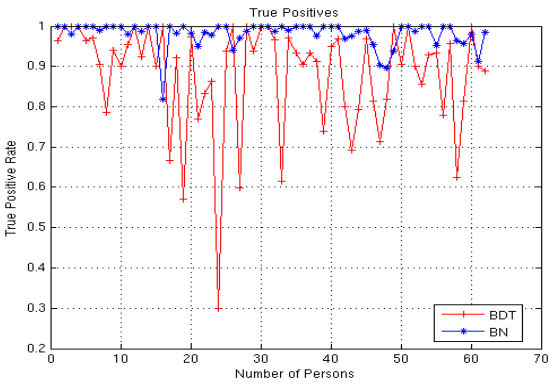
Table 1 Results of Our Approach

	Binary Decision Tree (BDT)	Bayesian Network (BN)
Face Recognition	93.65%	98.69%
Facial Expressions Recognition	98.30%*	92.63%
Gender Classification	97.39%	97.88%*

Table 2 Results from 3D model

	Binary Decision Tree (BDT)	Bayesian Network (BN)
Face Recognition	98.49%	96.42%
Facial Expressions Recognition	97.56%*	99.67%
Gender Classification	99.08%	99.62%*

*****Classifier is used in combination to AdaboostM1.



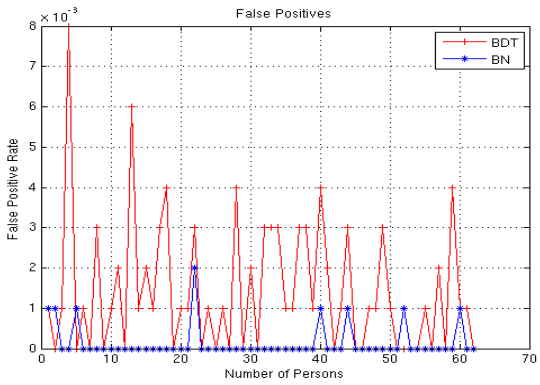
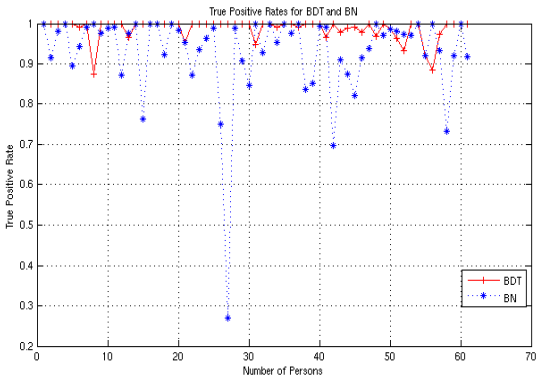


Figure 6. True positives and false positives for 2D model using two different classifiers.



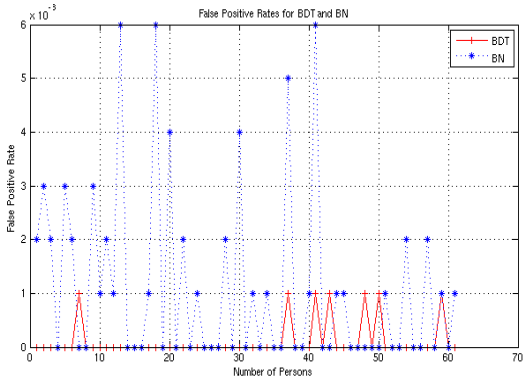


Figure 7. True positives and false positives for 3D model using two different classifiers.