Spatio-Temporal Multifeature for Facial Analysis

Zahid Riaz and Michael Beetz

Intelligent Autonomous Systems (IAS), Technical University of Munich,
Karlstr. 45, D-80333 Munich, Germany
{riaz,beetz}@in.tum.de

Abstract. Human faces are 3D complex objects consisting of geometrical and appearance variations. They exhibit local and global variations when observed over time. In our daily life communication, human faces are seen in actions conveying a set of information during interaction. Cognitive science explains that human brains are capable of extracting this set of information very efficiently resulting in a better interaction with others. Our goal is to extract a single feature set which represents multiple facial characteristics. This problem is addressed by the analysis of different feature components on facial classifications using a 3D surface model. We propose a unified framework which is capable to extract multiple information from the human faces and at the same time robust against rigid and non-rigid facial deformations. A single feature vector corresponding to a given image is representative of person's identity, facial expressions, gender and age estimation. This feature set is called spatio-temporal multifeature (STMF) extracted from image sequences. An STMF is configured with three different feature components which is tested thoroughly to evidence its validity. The experimental results from four different databases show that this feature set provides high accuracy and at the same time exhibits robustness. The results have been discussed comparatively with different approaches.

1 Introduction

In our daily life human faces are generally observed in action and convey a large set of information at a glance. On the other hand human brains have the ability to process this information very quickly and extract a distinct set of feature to classify various facial attributes. We build our concept on the fact from cognitive science study that facial features are processed holistically [1]. For example, instead of calculating local descriptor from facial geometry, color and corners, human brains process this information rather holistically. In general, we come across different people especially in social gatherings, meetings, markets and public places. With a first glance at their faces, we acquire a set of information about them even without interaction. This information contains gender, facial behavior, age estimation, ethnic origin and identity of the person if known before. This can be seen in Figure 1. Our goal is to find a single feature set which is representative of these facial attributes. We provide a feature set which

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 199–209, 2012.

generates high accuracy and at the same time is robust against currently outstanding challenges like varying head poses and facial expressions. This feature set is extracted using a 3D face model and mainly consists of spatio-temporal components of the faces. A 3D model is computed from a single image of a human face. We term the extracted feature set spatio-temporal multifeature (STMF) because it is composed of multiple facial parameters which include geometrical, appearance and deformation components. This benefits in controlling the spatial variation in the form of structure and texture whereas, motion pattern in the form of temporal features which jointly provides sufficient strength over different facial dynamics. The major contributions of this paper are; (i) to provide a unified feature set which is representative of facial expressions, gender, facial age and identity of a person, (ii) to show that the extracted feature set is robust against varying facial poses and expressions for face recognition applications, (iii) to comparatively study 2D and 3D face models showing that a 3D model outperforms 2D model, (iv) and finally 3D models are capable to remove perspective distortions in texturing a face which provide improved results. The results have been extracted on four different benchmark datasets in comparison to different state-of-the-art approaches.

The remaining part of this paper is divided in four sections. Section 2 discusses some of the related work in face image analysis. Section 3 explains problem statement in detail and configuration of our feature set. Section 4 describes detailed feature extraction approach. Finally, section 5 evaluates the calculated features on three different databases with various experimentations.

Fig. 1. With a glance at a face, humans are capable to classify different facial attributes. Our goal is to find a single feature set which is representative of these facial attributes.

2 Related Work

One of the major applications of human face image analysis is person identification. Over the efforts of a couple of decades, researchers are able to develop commercial systems for face recognition. Many industries are providing solution for facial biometrics [2]. Another important representation of the faces is the behavior of the persons, which is interpreted in the form of facial expressions. Other facial traits which do not directly represent person identity are categorized as soft-biometrics [3]. These include facial expressions, gender, age estimation and ethnicity. In the literature of face modeling, some useful face models are point distribution models (PDM), 3D morphable models, photorealistic models, deformable models and wireframe models [4]. Our proposed features consist of model parameters which fulfill a model's key requirements of dealing with nonrigidness, deformation and capability to synthesize novel views. 3D morphable models [5], which are available with different variants [6][7] are useful in synthesizing novel views however are not efficient and require a manual intervention in preprocessing step. These models are dense and finer than 2D active appearance models (AAM) [8] and suitable for facial synthesis and animation. We use rather a coarser model called Candide-III [9] which is defined with 184 triangular surfaces. This model is supported by action units and MPEG-4 animation units [10] which makes it useful for facial expressions analysis. Candide-III is a shape model but can be rendered with texture.

We emphasize on 3D modeling for the faces because of several benefits. One of the major advantages of model parameters is the better control over facial dynamics, structure and appearance. Model parameters have been varied by Vetter et al [5] to control expressions and gender. Our system solve similar problem but using very coarser model with normal camera images and achieve high accuracy, robustness as well as fully automatic [11]. Multifeature fusion approach has also been used by Riaz et. al. [12] however robustness of the system against facial poses and expressions is not studied. In [13] authors have tested several techniques toward model fitting and their performance for facial expressions on Cohn-Kanade database. We compare our results with their [13] results but using different model fitting methodology. Hadid et al [14] discuss manifold learning for gender classification and results are obtained on different databases. However, with model based approach better accuracy is reported since model parameters contain more detail. In this paper, we extend the approach presented in [15] and apply a similar system for person identification, facial expressions, gender and estimation about aging. In addition to that we study effect of perspective distortion on human faces as compared to conventional AAM.

3 Spatio-Temporal Multifeature (STMF)

In this section, we configure a feature set which is representative of multiple facial attributes. Currently available systems for facial classification lack the property of multifeature representation of the faces. A major reason is that the most of the research work suggests to isolate the sources of variations while focusing on a particular application. For example, in face recognition application, it is a general practice to normalize the faces in order to remove facial expressions variations to stabilize face recognition results against unwanted facial expressions. Hence the extracted features do not contain facial deformation. We address an idea to develop a unified feature set which contain information regarding face recognition, facial expressions, facial aging and gender classification.

For face recognition textural information plays a key role in features extraction [16][10] whereas facial expressions are mostly person independent and require motion and structural components [17]. Similarly facial structure and texture vary significantly between gender and age classes. On the basis of this knowledge and literature survey we categorize three major feature components

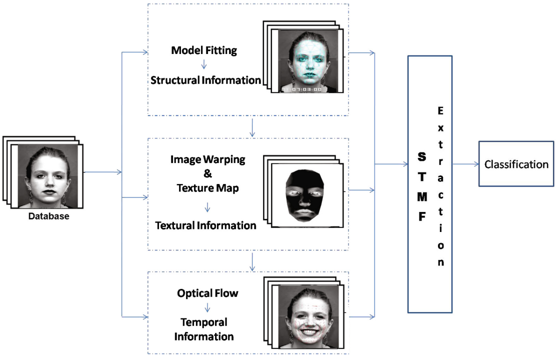


Fig. 2. Our Approach: Sequential flow toward feature extraction

Table 1. Contribution of different components to an STMF, p = primary contribution, s = secondary contribution, na = not applicable. This configuration is provided by a thorough literature survey. The experimental evidence of this configuration is provided in section 5.

	Identity	Expressions	Gender	Ethnicity	Age
Structural	p	p	p	p	p
Textural	p	s	p	p	p
Temporal	s	p	s	s	na

as primary and secondary contributor to five different facial classifications. Table 1 summarizes the significance of these constituents of the feature vector with their primary and secondary contribution toward the feature set formation. Since our feature set consists of all three kinds of information hence it can successfully represent facial identity, expressions, age, ethnicity and gender. The results are discussed in detail in section 5. Since in real world faces are seen from different views under varying poses and deformations therefore we use 3D modeling of the faces. However, any other approach can also be applied here with similar feature configuration. The approach presented in this paper is shown in Figure 2.

4 Features Extraction

In this section we explain our approach in different modules including shape model fitting, texture extraction and generating undistorted texture map and finally temporal parameters to finalize an STMF. The system initializes with a localization of a face in an image. We apply the algorithm by Viola and Jones [18] to detect and locate the face position within the image.

4.1 Structural Features

Structural features are obtained after fitting the model to the face image. For model fitting, local objective functions are calculated using haar-like features.

Face detector localizes the face region within a window. We use a state-of-the-art approach for model fitting using best objective functions. This approach is less prone to errors because of better quality of annotated images which are provided to the system for training. Further, this approach is less laborious because the objective function design is replaced with automated learning. For details we refer to [19][11]. The geometry of the model is controlled by a set of action units and animation units. Any shape $\mathbf{s}$ can be written as a sum of mean shape $\overline{\mathbf{s}}$ and a set of action units and shape units.

$$s(\alpha, \sigma) = \overline{s} + \phi_a \alpha + \phi_s \sigma \tag{1}$$

Where $\phi_\mathbf{a}$ is the matrix of action unit vectors and $\phi_\mathbf{s}$ is the matrix of shape vectors. Whereas α denotes action units parameters and σ denotes shape parameters [10]. Model deformation governs under facial action coding systems (FACS) principles [20]. The scaling, rotation and translation of the model is described by

$$s(\alpha, \sigma, \pi) = mRs(\alpha, \sigma) + t \tag{2}$$

Where $\mathbf{R}$ and $\mathbf{t}$ are rotation and translation matrices respectively, m is the scaling factor and π contains six pose parameters plus a scaling factor. By changing the model parameters, it is possible to generate some global rotations and translations.

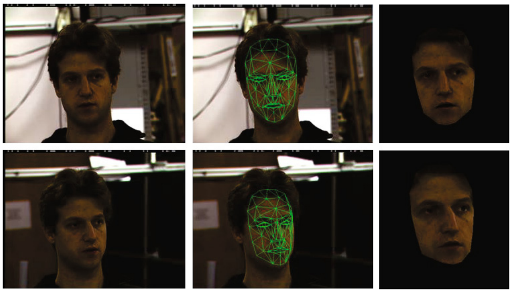


Fig. 3. Model fitting to two example images and texture projection on 3D surface after perspective correction. Images are taken from CMU-PIE database [21].

4.2 Textural Features

We take benefit of the 3D model to get texture projected to 3D surface and store undistorted texture in the form of texture map. The robustness of textural parameters depends upon the quality of the input texture image. Generally the effects of perspective distortions in the face images are ignored because the distance between the camera and the face is larger than the face size. We consider this effect because affine warping of the rendered triangle is not invariant to 3D rigid transformations. Since most of the triangles on the rotated faces and the face edges are tilted, therefore texture is heavily distorted in these triangles. In order to solve this problem we first apply perspective transformation followed by affine transformation to store texture in texture maps. The final transformation $\mathbf{M}$ is given by:

$$M = AK \left[R \; -Rt \right] \tag{3}$$

Where $\mathbf{K}$, $\mathbf{R}$ and $\mathbf{t}$ denotes the camera matrix, rotation matrix and translation vector respectively. Whereas $\mathbf{A}$ is affine transformation of extracted texture to texture map. For further details, refer [22].

Figure 4 shows the texture map for a single image along with a synthesized face with smiling action using this texture. The extracted texture coefficients are parametrized using PCA with mean vector $\mathbf{g_m}$ and matrix of eigenvectors $\mathbf{P_g}$ to obtain the parameter vector $\mathbf{b_g}$ [15].

$$\boldsymbol{g} = \boldsymbol{g_m} + \boldsymbol{P_g b_g} \tag{4}$$

Where $\mathbf{g}$ is the texture vector.

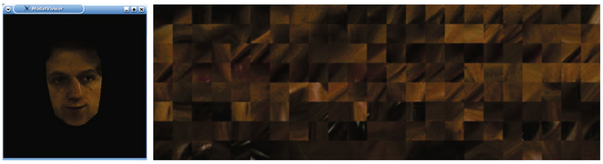


Fig. 4. Synthesized novel view from Figure 3 and its texture map

4.3 Temporal Features

Further, temporal features of the facial deformation are also calculated which consider motion over time. Local motion of feature points is observed using Lucas-Kanade pyramidal optical flow [23] on the model points. Figure 5 shows the motion patterns from image sequences in our database. We again use PCA over the velocity vectors obtained from the motion of feature points. If $\mathbf{t}$ is the velocity vector,

$$\boldsymbol{t} = \boldsymbol{t_m} + \boldsymbol{P_t b_t} \tag{5}$$

Where temporal parameters $\mathbf{b_t}$ are computed using matrix of eigenvectors $\mathbf{P_t}$ and mean velocity vectors $\mathbf{t_m}$.

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 additional temporal features. The overall feature vector becomes:

$$\boldsymbol{u} = (\boldsymbol{b_{s1}}, ..., \boldsymbol{b_{sm}}, \boldsymbol{b_{g1}}, ..., \boldsymbol{b_{gn}}, \boldsymbol{b_{t1}}, ..., \boldsymbol{b_{tp}}) \tag{6}$$

Where $\mathbf{b_s}$, $\mathbf{b_g}$ and $\mathbf{b_t}$ are shape, textural and temporal parameters respectively with m, n and p being the number of parameters retained from subspace in each

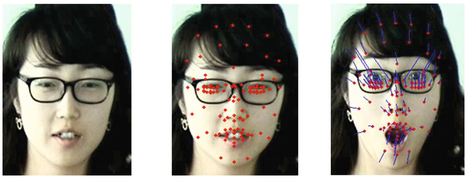


Fig. 5. Motion patterns from an image sequences. Initial image (left), landmark points using model fitting (middle), motion of the landmark points (right).

Table 2. Comparison of traditional AAM approach and rectified texture. The results are shown for textural parameters and combined structural and textural parameters. Face recognition on PIE dataset (above row), age estimation on FG-NET dataset (below row).

Database	**2D Texture parameters**	**Rectified Textural Parameters**	**AAM**	**3D Structural + Textural Parameters**
PIE	63.02%	**69.64%**	79.93%	**84.15%**
FG-NET	51.35%	**54.09%**	51.15%	**55.39%**

case. Equation 6 is called an *STMF vector*. We extract 85 structural features, approximately 200 textural features and 12 temporal features to form a combined feature vector for each image. All the subjects in the database are labeled for classification. All three individual components are normalized to one before fusion.

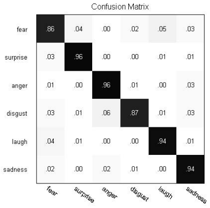
5 Experimentation

In order to validate the extracted features, we have used different subjects from four different databases called, CMU-PIE database [21], MMI database [24], Cohn-Kanade facial expressions database (CKFED) [25] and FG-NET database [26]. The texture extracted in each case is stored as a texture map after removing perspective distortion. During all experiments, we use two-third of the feature set for building the classifier model with 10-fold cross validation to avoid over fitting. The remaining feature set is used for testing purpose. We use all subjects from MMI, CKFED and FG-NET databases and partially use PIE database. We extract temporal features only from image sequences. For example, in case of FG-NET database no temporal information is available and hence we use only structural and textural components of the feature vector.

Since STMF set arises from different sources, so decision tree (DT) is applied for classification. However, other classifiers can also be applied here depending upon the application (Support Vector Machine (SVM) and Bayesian Networks (BN) were also used with comparable results during experimentation). We choose J48 decision tree algorithm for experimentation which uses tree pruning called subtree raising and recursively classifies until the last leave is pure. We use same

Table 3. Facial expressions recognition using different combinations of the feature sets. It can be seen that the optimal performance is obtained using all three components. These results conform to concept of feature configuration given in Table 1 (Left) Confusion matrix from our results (Right).

Feature type	**% Recog**
Shape + temporal	92.2%
Shape + Texture	83.3%
Shape + Texture + Temporal	96.4%


Confusion Matrix

	fear	surprise	anger	disgust	laugh	sadness
fear	.86	.04	.00	.02	.05	.03
surprise	.03	.96	.00	.00	.01	.01
anger	.01	.00	.96	.01	.00	.03
disgust	.03	.01	.06	.87	.01	.03
laugh	.04	.01	.00	.00	.94	.01
sadness	.02	.00	.02	.01	.00	.94

configuration for all classifiers trained during the experiments. The parameters used in decision tree are: confidence factor C = 0.25, with two minimum number of instances per leaf and C4.5 approach for reduced error-pruning [27].

5.1 Face Recognition

Face recognition experiments are performed on the aforementioned databases which contain images with six different facial expressions, frontal and profile faces and normal talking faces. We use all subjects from MMI and CKFED and 34 subjects with approximately 5000 images from two different sessions of the PIE database. The results of face recognition in the presence of varying poses are performed on PIE dataset and are shown in Table 2 in comparison to AAM. These results show that the removal of perspective distortions improves the classification rate. The classification rates for CKFED and MMI have been 99.59% and 99.29% respectively.

Table 4. Comparison of gender classification in comparison to different approaches [14]

Approach	Classification rate
Pixels + SVM + Fusion	88.5%
LBP + SVM + Fusion	92.1%
VLBP + SVM	84.5%
EVLBP + AdaBoost	84.6%
STMF + DT	**94.8%**

5.2 Facial Expressions Recognition

Facial expressions recognition is performed on CKFED with six universal facial expressions: anger, disgust, fear, laugh, sadness and surprise. We exclude neutral expression during the experiments, however a neutral expression can also be considered as a seventh expression during classification. An STMF feature set with three constituents (structural, textural and temporal) are used for experiments. The results are shown in Table 3 with different feature components along with the confusion matrix from our experiments. It can be seen that the results on left column of Table 3 coincide with our initial concept presented in Table 1 which shows the validity of our feature configuration.

5.3 Gender and Age Classification

We use same set of STMF for gender classification on CKFED. The results are shown in Table 4 in comparison to other approaches. (The results in Table 4 are not reproduced and use 10-fold cross validation as that of 5-fold cross validation described in [14]). The classification results may vary with choice of different classifier. For instance, we obtained 96.8% accuracy using random forests classifier instead of decision trees. This classification rate is comparable to the best reported result in [14] which is 97.2%. In order to verify our feature configuration concept, we further study age classification from all subjects of FG-NET database. This database consists of 1002 images of 62 subjects with age ranging from $0 - 69$ years. We divide the database in seven groups with 10 years band. The results are shown in Table 2 in comparison to AAM and with and without structural components.

6 Conclusions and Future Work

In this paper we have experimented 3D model based STMF extraction approach. Our aim is to find the feature configuration which directly reflects human perception about the faces. Different researchers have used model based approaches and found different solutions to biometric and soft-biometric traits using various features. However, concrete research under such scenarios is still required. Moveover model fitting and efficiency of the algorithm are challenges. So far we have studied four different traits of human faces called person identity, gender, facial expressions and aging however it can further be applied to ethnicity classification. Our future goal is to consider these traits on a large variability by further enhancing our feature set for interactive systems.

Acknowledgments. This research work is funded by German Academic Exchange Service (DAAD) under the project ID-54378355 on *Context Aware Perception for Assistive Intelligent Robots*, between German and Pakistani universities (2011-2013).

References

1. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: Face recognition by humans: Nineteen results all computer vision researchers should know about. Proceedings of IEEE 94, 1948–1962 (2006)
2. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey (December 2003)

3. Jain, A.K., Dass, S.C., Nandakumar, K.: Soft Biometric Traits for Personal Recognition Systems. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 731–738. Springer, Heidelberg (2004)
4. Abate, A., Nappi, M., Riccio, D., Sabatino, G.: 2d and 3d face recognition: A survey. Pattern Recognition Letters 28, 1885–1906 (2007)
5. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9), 1063–1074 (2003)
6. Huang, J., Heisele, B., Blanz, V.: Component-based Face Recognition with 3D Morphable Models. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 27–34. Springer, Heidelberg (2003)
7. Romdhani, S.: Face Image Analysis using a Multiple Feature Fitting Strategy. PhD thesis, Computer Science Department, University of Basel, Basel Switzerland (2005)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
9. Ahlberg, J.: An experiment on 3D face model adaptation using the active appearance algorithm. Image Coding Group, Deptt of Electric Engineering, Linköping University (2001)
10. Ahlberg, J., Dornaika, F.: Parametric Face Modeling and Tracking. In: Li, S.Z., Jain, A.K. (eds.) Handbook of Face Recognition. Springer (2005)
11. Mayer, C., Wimmer, M., Stulp, F., Riaz, Z., Roth, A., Eggers, M., Radig, B.: A real time system for model-based interpretation of the dynamics of facial expressions. In: Proc. of the International Conference on Automatic Face and Gesture Recognition (FGR 2008), Amsterdam, Netherlands (September 2008)
12. Riaz, Z., Mayer, C., Beetz, M., Radig, B.: Model Based Analysis of Face Images for Facial Feature Extraction. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 99–106. Springer, Heidelberg (2009)
13. Asthana, A., Saragih, J., Wagner, M., Goecke, R.: Evaluating aam fitting methods for facial expression recognition. In: Affetive Computing and Intelligent Interaction, vol. 1, pp. 598–605 (September 2009)
14. Hadid, A., Pietikäinen, M.: Manifold Learning for Gender Classification from Face Sequences. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 82–91. Springer, Heidelberg (2009)
15. Riaz, Z., Gedikli, S., Beetz, M.: Spatio-temporal facial features for hri scenarios. In: CRV. IEEE (2011)
16. Zhao, W., Chellappa, R.: Face Processing: Advanced Modeling and Methods. Elsevier Inc. (2005)
17. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36(1), 259–275 (2003)
18. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57, 137–154 (2004)
19. Wimmer, M., Stulp, F., Tschechne, S., Radig, B.: Learning robust objective functions for model fitting in image understanding applications. In: Proceedings of the 17th British Machine Vision Conference, pp. 1159–1168 (2006)
20. Ekman, P., Friesen, W.: The facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press (1978)

21. Terence, S., Baker, S., Bsat, M.: The cmu pose, illumination, and expression (pie) database. In: FGR 2002: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, p. 53. IEEE Computer Society Press, Washington, DC (2002)
22. Riaz, Z., Beetz, M.: On the effect of perspective distortion on face recognition. In: 12th International Conference on Vision Applications (VISAPP) (February 2012)
23. Bouguet, J.-Y.: Pyramidal implementation of the lucas kanade feature tracker (2000)
24. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis (July 2005)
25. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
26. http://www.fgnet.rsunit.com/
27. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)