Geometry vs. Appearance for Discriminating between Posed and Spontaneous Emotions

Geometry vs. Appearance for Discriminating between Posed and Spontaneous Emotions

Ligang Zhang, Dian Tjondronegoro and Vinod Chandran

Queensland University of Technology, 2 George Street, Brisbane, 4000, Australia
ligzhang@gmail.com, {dian, v.chandran}@qut.edu.au

Abstract. Spontaneous facial expressions differ from posed ones in appearance, timing and accompanying head movements. Still images cannot provide timing or head movement information directly. However, indirectly the distances between key points on a face extracted from a still image using active shape models can capture some movement and pose changes. This information is superposed on information about non-rigid facial movement that is also part of the expression. Does geometric information improve the discrimination between spontaneous and posed facial expressions arising from discrete emotions? We investigate the performance of a machine vision system for discrimination between posed and spontaneous versions of six basic emotions that uses SIFT appearance based features and FAP geometric features. Experimental results on the NVIE database demonstrate that fusion of geometric information leads only to marginal improvement over appearance features. Using fusion features, surprise is the easiest emotion (83.4% accuracy) to be distinguished, while disgust is the most difficult (76.1%). Our results find different important facial regions between discriminating posed versus spontaneous version of one emotion and classifying the same emotion versus other emotions. The distribution of the selected SIFT features shows that mouth is more important for sadness, while nose is more important for surprise, however, both the nose and mouth are important for disgust, fear, and happiness. Eyebrows, eyes, nose and mouth are important for anger.

Keywords: Facial expression, posed, spontaneous, SIFT, FAP.

1 Introduction

A machine vision system that can accurately discriminate spontaneous from posed expressions can be useful in ways similar to a polygraph. Spontaneous expressions are difficult to distinguish from posed ones and differ in subtle ways in appearance, timing and accompanying head movement [1]. In general, facial expression recognition (FER) can be reliably performed from still images with far less complexity. However, the timing and head movement can only be extracted from video, not from still images. The distances between key points on a face extracted from a still image can indirectly capture some movement and pose changes using active shape models (ASM). This information is superposed on information about non-rigid facial movement that is also part of the expression.

Relatively little research has been conducted on machine discrimination between posed and spontaneous facial expressions. These efforts mainly focus on smile [1], [2], [3], eyebrow action [4] and pain [5], [6]. For smile and eyebrow action, nearly all known approaches are designed based on the movements of facial points. Hamdi *et al.* [2] used eyelid movements and reported 85% and 91% accuracy in discriminating between posed and spontaneous smiles on the BBC and Cohn-Kanade databases, respectively. Michel *et al.* [3] proposed a multimodal system to discern posed from spontaneous smiles by fusing a set of temporal attributes of tracked points of face, head and body. 94% accuracy was the best result, obtained with late fusion of all modalities. Michel *et al.* [4] also proposed to use the temporal dynamics of facial points to distinguish between posed and spontaneous brow actions, and attained a 90.7% classification rate. Littlewort *et al.* [5] employed a two-stage system to differentiate faked pain from real pain: a detection stage for 20 facial actions using Gabor features and a SVM classification stage, achieving 88% accuracy. Bartlett *et al.* [6] reported 72% classification accuracy of posed versus spontaneous pain through Gabor feature based facial action detection. Other facial expressions, such as anger, disgust, fear, sadness, and surprise have not been fully investigated in this context.

Appearance features (e.g. SIFT and Gabor) are more suitable for capturing the subtle changes of the face; while geometric features (e.g. distance between landmarks) are more capable of representing shape and location information of facial components. Although the fusion of appearance and geometry leads to significant performance improvements on basic facial expression classification [7], it remains unclear whether such improvements are possible for posed versus spontaneous emotion discrimination as well. This paper addresses these areas.

An automatic system to distinguish posed from spontaneous versions of six basic emotions using appearance (SIFT) and geometric (FAP) features is adopted to investigate recognition performance. Feature selection is performed using minimal redundancy maximal relevance (mRMR) and classification using a support vector machine (SVM). Scale-invariant feature transform (SIFT) has been shown a better recognition performance of facial expressions in previous work [8] than other appearance features, including LBP and HOG, while the distances defined based on facial animation parameters (FAPs) have also been demonstrated as a sparse, compact, yet information-rich representation of the facial shape [9]. But they have not been combined yet for discriminating posed versus spontaneous emotions. Appearance features are extracted locally around key points while distances between key points are used for the geometric feature set. They are therefore expected to not contain overlapping information. Intuitively, geometry is not expected to result in significant improvement in performance in this context because no temporal information is captured. This paper will compare the relative importance of the two types of features and those extracted from different regions on the face for each of six emotions.

The rest of the paper is organized as follows. Section 2 presents the evaluation system. Section 3 gives the experimental results. Conclusions are drawn in Section 4.

2 Evaluation Framework

Fig. 1 shows the framework of the evaluation system. From an input image, the face region is detected using the widely used Viola-Jones detector, and 68 facial fiducial points are detected using a well-trained active shape model (ASM). SIFT descriptors are extracted around each of 53 interior facial points. Feature vectors from all points are concatenated into a single vector representing *appearance* features. A subset of the most discriminative appearance features is selected using the mRMR algorithm. *Geometric* features composed of 43 distances defined using an active shape model (ASM) and FAPs are also extracted. The normalized *appearance* feature subset and *geometric* features can be used alone or combined through a feature-level fusion. A SVM with a radial basis function (RBF) kernel is used as the classifier for discriminating between posed and spontaneous versions of six basic emotions - anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU).

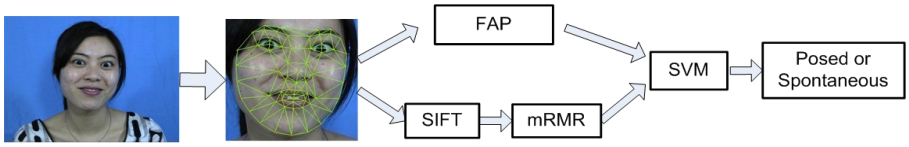
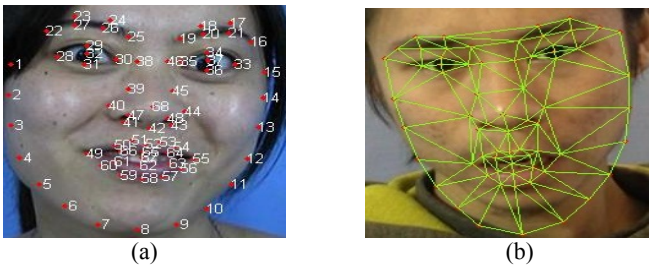


Fig. 1. Framework of the evaluation system.

2.1 Face and Fiducial Point Detection

Once the face region is detected by the Viola-Jones detector, an ASM [10] is used to detect the fiducial points. To train the ASM, we collected 100 images from the internet with different natural emotions and different face poses ranging from -20 to 20 degrees. Then 68 fiducial points as shown in Fig.2a are manually annotated with x and y locations. The trained ASM is expected to work well on faces with normal face movements. It has been observed that the points on the face boundary (Index from 1 to 15 in Fig.2a) are not always accurately detected due to facial shape changes between subjects and facial movements (as shown in Fig.2b). Further, the regions around these points contain background information and do not provide reliable features. Therefore, only 53 interior points (Index from 16 to 68 in Fig.2a) are used to extract SIFT features.



(a) (b)

Fig. 2. (a) 68 fiducial points for training ASM and (b) detection results with inaccurate boundary points.

2.2 SIFT Feature Extraction

SIFT [11] provides distinctive invariant features suitable for detecting salient key points and describing local appearance. The SIFT is known to be invariant to image scale and rotation, and robust across a substantial range of affine distortion, changes in 3D viewpoint, noise and illumination. SIFT features extracted around a small set of facial landmarks have been applied to describe local characteristics of the face, yielding promising results [12]. SIFT features around a number of points also helps to achieve a degree of tolerance to face movements and pose changes.

Following the settings in [12], the SIFT descriptor is computed from the gradient vector histograms of the pixels in a 4×4 patch around each point of 53 interior points. Instead of setting a fixed orientation, we let the program compute the 8 possible gradient orientations. Therefore, each SIFT descriptor contains a total of 128 elements. By computing one such descriptor at each point, we obtain a final feature vector with 6,784 elements.

2.3 Geometric Feature Extraction

Facial animation parameters (FAPs) [13] are defined in the ISO MPEG-4 standard (part 2, visual) to allow the animation of synthetic face models. FAPs contain 68 parameters that are either high level parameters describing visemes and expressions, or low level parameters describing displacements of the single points of the face. Therefore, FAPs can provide a concise representation of the evolution of the expression of the face and can represent a complete set of basic facial actions. Furthermore, FAPs also can handle arbitrary faces through the use of FAP units (FAPUs), which are defined as the fractions of distances between key points.

Geometric features include 43 distances between the 53 interior points. As listed in Table 1, these distances are calculated based on FAPs to allow the animation of face shape changes. Compared with facial movement vectors in multi-frames, distance features have the merit of being robust to pose changes, and do not require

Table 1. Distances between facial points defined by FAPs.

No.	Distance	No.	Distance	No.	Distance	NO.	Distance
3	Dy(52,58)	19	Dy(29,32)	33*	Dy(32,27)	55	Dy(50,42)
4	Dy(65,42)	20	Dy(34,37)	34*	Dy(37,17)	56	Dy(54,42)
5	Dy(62,42)	21	Dy(31,32)	34*	Dy(37,18)	57	Dy(60,42)
6	Dx(49,42)	22	Dy(36,37)	34*	Dy(37,20)	58	Dy(56,42)
7	Dx(55,42)	29	Dy(29,31)	34*	Dy(37,21)	61*	Dx(30,40)
8	Dy(66,42)	30	Dy(34,36)	35	Dy(28,22)	61*	Dx(30,39)
9	Dy(64,42)	31	Dy(30,25)	36	Dy(33,16)	62*	Dx(35,44)
10	Dy(61,42)	32	Dy(35,19)	37	Dx(30,25)	62*	Dx(35,45)
11	Dy(63,42)	33*	Dy(32,23)	38	Dx(35,19)	63	Dy(35,68)
12	Dy(49,42)	33*	Dy(32,24)	51	Dy(52,42)	64	Dx(35,68)
13	Dy(55,42)	33*	Dy(32,26)	52	Dy(58,42)	-	-

Note: Dx(M,N) and Dy(M,N) indicate the distances between two points indexed M and N in the horizontal and vertical directions respectively. M and N are based on the 53 interior points in Fig. 2a.

compensation for face movements. The distances defined based on FAPs have been demonstrated as a sparse, compact, yet information-rich representation of the facial shape [9]. Therefore, they are suitable for the proposed system. To allow for variations between faces, FAPUs are defined as the fractions of distances between key points to scale FAPs (i.e. 43 distances).

2.4 Discriminative Texture Feature Selection

We use the minimal redundancy maximal relevance criterion (mRMR) [14] algorithm to select a subset of the most discriminative features from the extracted SIFT features. The mRMR selects a subset of features that jointly have the largest dependency on the ground truth class and the least redundancy among the features, according to following equation:

$$\max_{x_j \in X - S_{m-1}} [I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i)] \tag{1}$$

Where x_i and x_j are the i^{th} and j^{th} features of the feature set X; I(x,y) is the mutual information of two variables x and y; c are the class labels. Suppose S_{m-1} is the already selected feature set with m-1 features, the task is to select the m^{th} feature from the set {X- S_{m-1}}.

When using mRMR, a discretization of the continuous inputs was recommended [14]. This paper obtains the discrete feature $\overline{D}_k$ of a continuous feature D_k based on the mean value μ_k and the standard deviation σ_k of all features:

$$\overline{D}_k = \begin{cases} -2 & if \quad D_k < \mu_k - \sigma \cdot \sigma_k \\ 0 & if \quad \mu_k - \sigma \cdot \sigma_k \le D_k \le \mu_k + \sigma \cdot \sigma_k \\ 2 & if \quad D_k > \mu_k + \sigma \cdot \sigma_k \end{cases} \tag{2}$$

Where σ is set to 0.5.

3 Experiments

3.1 Database

The natural visible and infrared facial expression (NVIE) database [16] is a newly developed comprehensive platform for both spontaneous and posed facial expression analysis. The spontaneous expressions are induced by film clips deliberately selected from the internet, while the posed ones are obtained by asking the subjects to perform a series of expressions. There are a total of 215 healthy students (157 males and 58 females), ranging in age from 17 to 31. Among them, 105, 111, 112 subjects participated in the spontaneous database under front, left and right illumination respectively, and 108 subjects participated in the posed database. Both spontaneous and posed images with peak emotions are labeled with six basic emotions.

(a) (b)

Fig. 3. (a) Image samples excluded from the experiment because of inaccurate detection results by ASM. (b) Images samples of posed versus spontaneous emotions.

Table 2. Distribution of the used images over six emotions.

	AN	DI	FE	HA	SA	SU
Posed	593	578	609	607	581	604
Spontaneous	229	266	211	315	236	215

In this paper, all posed peak visible images are used, while only spontaneous visible images with final evaluated annotations are used. Note that a part of spontaneous images have not been provided with final annotations by the time of writing this paper. After removing those failed during face and facial point detection, we get 3,572 posed and 1,472 spontaneous images. Fig. 3a shows samples of removed images due to inaccurate ASM detection results of facial points. As can be seen, the mouth region and the face boundary are less likely to be accurately detected by the ASM when big out-of-plane rotations of the face occur (i.e. pitch and yaw). Fig. 3b demonstrates samples of posed versus spontaneous emotions and Table 2 shows the distribution of the images over six emotions.

3.2 Classification Performance

We conducted subject-independent tests to obtain an average classification result over 10 cross-validations. In each cross-validation, images of 10% subjects are randomly selected for testing and the images of 90% subjects left are for training. The process repeats 10 times to obtain average classification accuracy. Note that the emotional labels of each of the six basic emotions are assumed to be known before classifying posed versus spontaneous emotions using a SVM.

Fig. 4 shows the accuracy of posed versus spontaneous classification of six emotions. As anticipated, fusion of SIFT and FAP features only leads to a marginally higher overall performance than using SIFT features only, for all emotions except for disgust. For disgust, inclusion of FAP features in fact leads to a lower performance than using SIFT features alone. The use of FAP features does not improve the performance, and this may be because FAP based distances have a limited capacity to capture the temporal information (e.g. movements) of facial expressions, while discrimination between posed and spontaneous emotions largely depends on such information as shown in previous studies [1], [17], [18]. The results agree with the claim in [18] that high-abstraction features extracted from video segments can capture more general physical phenomena than low-abstraction features in one frame. In addition, posed and spontaneous emotions in static images are more likely to have similar geometric distances, and their differences are mainly conveyed by subtle

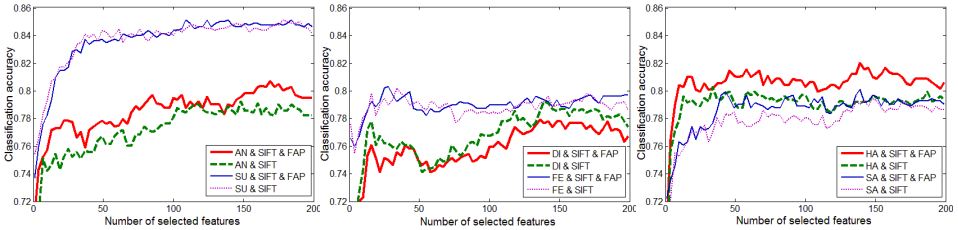


Fig. 4. Classification accuracy for posed versus spontaneous versions of six basic emotions on the NVIE database. As can be seen, fusion of SIFT and FAP only leads to marginally higher overall performance than using SIFT features only. The accuracy of using FAP features alone can be observed from the third row in Table 3. For disgust, fusion is in fact catastrophic.

Table 3. Posed/Spontaneous classification accuracy (%) + one standard deviation.

	AN	DI	FE	HA	SA	SU
SIFT+FAP	77.2±6.3	76.1±6.1	79.7±6.2	80.5±6.1	79.7±5.1	83.4±5.2
SIFT	75.6±7.1	76.1±6.9	79.6±7.3	79.3±7.7	77.4±7.6	83.9±4.0
FAP	71.2 ±4.9	69.7±6.0	76.2±4.0	65.6±5.8	68.7±4.8	73.3±6.3

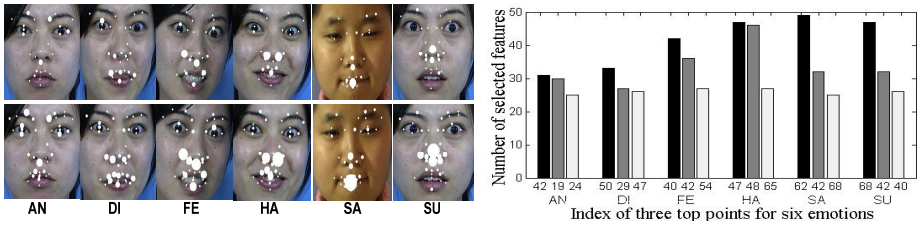
appearance features. It should be noted that a subset of facial points located by ASM without enough precision could also introduce noise to FAP features.

The adopted system using SIFT+FAP or SIFT features obtains an accuracy of more than 74% for all emotions. Among the six emotions, surprise and happiness are the two easiest ones to be distinguished as posed or spontaneous, whereas disgust is the most difficult one. The results are similar to those obtained on recognition of six basic emotions, in which surprise and happiness are often the easiest emotions to recognize, while disgust is one of the most difficult ones. It also can be observed from Fig. 4 that anger, fear and sadness have a similar performance.

Table 3 demonstrates the classification accuracy for posed versus spontaneous emotions based on 40 SIFT and 43 FAP features. Using SIFT+FAP features, the employed system obtains the highest accuracy of 83.4% when testing on surprise, and the lowest accuracy of 76.1% when testing on disgust. SIFT+FAP and SIFT have a similar performance for most of the six emotions, and they both outperform using FAP features alone. For anger, disgust, fear, and surprise, the performances of SIFT+FAP, SIFT and FAP are not statistically significantly different as observed when one standard deviation intervals are noted in Table 3. For happiness and sadness, the performance of SIFT+FAP is also similar to SIFT, but is one standard deviation more significant than FAP. This result again implies that geometric features play a less important role than appearance features on posed versus spontaneous emotion discrimination, and including geometric features leads to little performance improvements compared with using appearance features alone.

3.3 Feature Importance Comparison

To investigate the importance of different points in their contribution to discrimination, we display the distribution of the selected SIFT features over the 53



(a) (b)

Fig. 5. (a) Distribution of selected SIFT features over 53 points. The two rows contain the top 40 and top 90 features selected by the mRMR in each of 10 subject-independent tests. Emotion class labels are given below. (b) The three top facial points selected for each of the six emotions. Index numbers are mapped to face locations in Fig. 2a.

interior points as shown in Fig.5a. If more features are selected from a given facial point, that point may be considered more important and is marked by a bigger white circle. It should be noted that similar distributions have also been observed using different subsets of the NVIE data.

We can observe that different facial elements play different roles in distinguishing different posed versus spontaneous emotions. The mouth appears to be more important for sadness, the nose region is more important for surprise, while both the nose and mouth are important for disgust, fear, and happiness. The eyebrows, eyes, nose and mouth all play a significant role for anger. The results are contrary to the findings in discriminating six emotions in thermal images [16], where the mouth region has the smallest impacts on all six emotions, and the nose has little impacts on sadness, surprise, fear, and happiness. However, one common point between our work and [16] is that the nose plays a significant role in classifying anger versus other emotions, and discriminating posed versus spontaneous anger. Similar importance of nose is also found for disgust. The important role of the nose in the evaluated framework contrasts with the common understanding that the nose is the relative expression-invariant facial region. Therefore, there appears to exist different important facial regions between discriminating posed versus spontaneous version of one emotion and classifying the same emotion versus other emotions. This is within our expectation as each emotion has its own discriminative facial regions when classifying it versus other emotions. However, discriminating posed versus spontaneous versions of the same emotion needs to depend on information in other regions.

From Fig. 5a, we also can see that feature points on the eyebrows and eyes seem to provide few of the top 40 or top 90 features, for most of the emotions. This is probably due to the fact that about a half of the faces in the NVIE database have glasses, which occlude the useful information in the eyebrows and eyes. In addition, feature points on the mouth also have different distributions for different emotions. For instance, the points focus on the corner lip for fear, the top lip for happiness, and the middle lip for sadness.

Fig.5b gives the three top facial fiducial points that contain the largest number of the selected features for each emotion. As can be seen, most of these points for six emotions are distributed on nose and the points indexed 42, 40, 47, 68 are shared by different emotions (e.g. the point 40 is shared by fear and surprise). Compared with

the top points for other emotions, the point 62 for sadness and the point 68 for surprise take a much larger proportion of the selected features.

4 Conclusions

A machine vision system for distinguishing posed and spontaneous versions of six basic emotions in static images is used to compare the performance of SIFT features from ASM based fiducial points, FAP distance features and their fusion. Experimental results show that appearance features play a significantly more important role than geometric features on posed versus spontaneous emotion discrimination, and fusion of geometric information leads only to marginal improvement over SIFT appearance features. This is owing to the fact that temporal information is not available in the geometric representation of still images. Among six emotions, surprise is the easiest emotion (83.4% accuracy) to be classified as posed or spontaneous, while disgust is the most difficult one (76.1%) using SIFT+FAP features. Our results find that there are different important facial regions between discriminating posed versus spontaneous version of one emotion and classifying the same emotion versus other emotions. In terms of providing the most relevant features for classification between posed and spontaneous emotions, the mouth is more important for sadness, the nose is more important for surprise, while both the nose and mouth are important for disgust, fear, happiness, and the eyebrows, eyes, nose, mouth are all important for anger. A significant proportion of the SIFT features selected by the mRMR for six emotions are distributed on the points in the nose region. Our future work will test the performance fusing SIFT features with temporal geometric features in video, and explore real world applications.

References

1. Cohn, J., Schmidt, K.: The timing of facial motion in posed and spontaneous smiles. International Journal of Wavelets, Multiresolution and Information Processing 2 (2004) 1-12
2. Hamdi, D., Roberto, V., Albert Ali, S., Theo, G.: Eyes do not lie: spontaneous versus posed smiles. Proceedings of the international conference on Multimedia. ACM, Firenze, Italy (2010) 703-706
3. Michel, F.V., Hatice, G., Maja, P.: How to distinguish posed from spontaneous smiles using geometric features. Proceedings of the 9th international conference on Multimodal interfaces. ACM, Nagoya, Aichi, Japan (2007) 38-45
4. Michel, F.V., Maja, P., Zara, A., Jeffrey, F.C.: Spontaneous vs. posed facial behavior: automatic analysis of brow actions. Proceedings of the 8th international conference on Multimodal interfaces. ACM, Banff, Alberta, Canada (2006) 162-170
5. Littlewort, G.C., Bartlett, M.S., Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image and Vision Computing 27 (2009) 1797-1803
6. Bartlett, M., Littlewort, G., Vural, E., Lee, K., Cetin, M., Ercil, A., Movellan, J.: Data Mining Spontaneous Facial Behavior with Automatic Expression Coding. Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Vol. 5042. Springer Berlin / Heidelberg (2008) 1-20

7. Mingli, S., Dacheng, T., Zicheng, L., Xuelong, L., Mengchu, Z.: Image Ratio Features for Facial Expression Recognition Application. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 40 (2010) 779-788
8. Yuxiao, H., Zhihong, Z., Lijun, Y., Xiaozhou, W., Xi, Z., Huang, T.S.: Multi-view facial expression recognition. Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on (2008) 1-6
9. Hao, T., Huang, T.S.: 3D facial expression recognition based on automatically selected features. Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on (2008) 1-8
10. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models-Their Training and Application. Comput. Vis. Image Underst. 61 (1995) 38-59
11. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60 (2004) 91-110
12. Berretti, S., Bimbo, A.D., Pala, P., Amor, B.B., Daoudi, M.: A Set of Selected SIFT Features for 3D Facial Expression Recognition. Pattern Recognition (ICPR), 2010 20th International Conference on (2010) 4125-4128
13. Pandzic, I.S., Forchheimer, R.: MPEG-4 facial animation: the standard, implementation and applications. Wiley (2002)
14. Hanchuan, P., Fuhui, L., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (2005) 1226-1238
15. Lajevardi, S., Hussain, Z.: Automatic facial expression recognition: feature extraction and selection. Signal, Image and Video Processing (2011) 1-11
16. Shangfei, W., Zhilei, L., Siliang, L., Yanpeng, L., Guobing, W., Peng, P., Fei, C., Xufa, W.: A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. Multimedia, IEEE Transactions on 12 (2010) 682-691
17. Schmidt, K., Ambadar, Z., Cohn, J., Reed, L.: Movement Differences between Deliberate and Spontaneous Facial Expressions: Zygomaticus Major Action in Smiling. Journal of Nonverbal Behavior 30 (2006) 37-52
18. Schmidt, K., Bhattacharya, S., Denlinger, R.: Comparison of Deliberate and Spontaneous Facial Movement in Smiles and Eyebrow Raises. Journal of Nonverbal Behavior 33 (2009) 35-45