Various Fusion Schemes to Recognize Simulated and Spontaneous Emotions

Sonia Gharsalli[1], Hélène Laurent[2], Bruno Emile[1] and Xavier Desquesnes[1]
[1]*Univ. Orléans, INSA CVL,*
PRISME EA 4229, Bourges, France
[2]*on secondment from INSA CVL, Univ. Orléans,*
PRISME EA 4229, Bourges, France
to the Rector of the Academy of Strasbourg, Strasbourg, France

Keywords: Facial Emotion Recognition, Posed Expression, Spontaneous Expression, Early Fusion, Late Fusion, SVM, FEEDTUM Database, CK+ Database.

Abstract: This paper investigates the performance of combining geometric features and appearance features with various fusion strategies in a facial emotion recognition application. Geometric features are extracted by a distance-based method; appearance features are extracted by a set of Gabor filters. Various fusion methods are proposed from two principal classes namely early fusion and late fusion. The former combines features in the feature space, the latter fuses both feature types in the decision space by a statistical rule or a classification method. Distance-based method, Gabor method and hybrid methods are evaluated on simulated (CK+) and spontaneous (FEEDTUM) databases. The comparison between methods shows that late fusion methods have better recognition rates than the early fusion method. Moreover, late fusion methods based on statistical rules perform better than the other hybrid methods for simulated emotion recognition. However in the recognition of spontaneous emotions, the statistical-based methods improve the recognition of positive emotions, while the classification-based method slightly enhances sadness and disgust recognition. A comparison with hybrid methods from the literature is also made.

1 INTRODUCTION

Automatic facial emotion recognition is a challenging topic in machine vision research. It has made many achievements in the last years in various applications (human/machine interaction, psychiatry, behavioural science, educational software, animation...).

Automatic facial emotion recognition methods can be distinguished in two main classes: geometric methods and appearance-based methods. Geometric methods detect face components shapes and positions. Feature points tracking and face motion trackers are the mostly used geometric techniques to capture expression of emotions from image sequences. Abdat et al (Abdat et al., 2011) represent each facial muscle motion by distance variation between pair of feature points. To recognize the six basic facial emotions and a set of Facial Action Units (FAU), Kotsia et al (Kotsia and Pitas, 2007) compute the displacements of some selected Candide nodes from the first frame to the greatest facial expression intensity frame. On the other hand, appearance-based methods extract facial texture changes such as wrinkles and furrows. These methods use various techniques to capture the skin texture changes such as Gabor wavelets (Bartlett et al., 2003), Local Binary Patterns (LBP) (Shan et al., 2009), optical flow (Anderson and McOwan, 2006).

Both geometric methods and appearance-based methods have some specific weaknesses. Kotsia et al (Kotsia et al., 2008b) report that the use of only texture information can lead to confusion between anger and fear emotions. However, the lack of texture information can lead to the misclassification of subtle facial movements. The combination between these two classes could then allow to achieve better results. Fasel et al (Fasel et al., 2002) explain that having an hybrid method can be of great interest, if the individual approaches produce very different error patterns.

The choice of the appropriate fusion scheme can also impact the results. The fusion of information is generally performed at two levels: feature level and decision level. For emotion recognition applications these two levels are highlighted when various modalities are combined such as: speech and facial

Gharsalli S., Laurent H., Emile B. and Desquesnes X..
Various Fusion Schemes to Recognize Simulated and Spontaneous Emotions.
DOI: 10.5220/0005312804240431
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 424-431
ISBN: 978-989-758-090-1

expressions (Busso et al., 2004), face and body gestures (Gunes and Piccardi, 2005). For facial expression recognition applications, the combination of different features is generally done by feature level fusion. Kotsia et al (Kotsia et al., 2008b) extract the appearance features by the Discriminant Non-negative Matrix Factorization (DNMF) methods. Besides, the shape is computed by the deformed Candide grid. An early fusion method is applied to combine between both descriptors. The same fusion scheme is applied by Zhang et al (Zhang et al., 2012) and Chen et al (Chen et al., 2012) to obtain robust combined features to recognize facial expressions. The geometric features are computed through distance-based method in (Zhang et al., 2012) and displacement-based method in (Chen et al., 2012). Both methods use in addition local texture information. Wan et al (Wan and Aggarwal, 2013) learn a distance metric structure from combined features. A feature level fusion is applied with different weights between texture and geometric features.

In this paper, various fusion strategies (early fusion, fusion by statistical rules and fusion by classification method) are studied and their robustness in the recognition of posed and spontaneous facial expressions is analysed.

The paper is organised as follows: description of the features extraction and the fusion strategies is proposed in the next section, followed by the presentation of the considered databases in section 3. Section 4 reports the experimental results on the CK+ database (Lucey et al., 2010) and the FEEDTUM database (Wallhoff, 2006). A discussion is also presented there. Conclusion and prospects are given in section 5.

2 METHODS DESCRIPTION

Emotion recognition systems are based on three steps: face detection, features extraction and features classification. In our work, we chose for real-time face detector an adapted version of Viola&Jones method (Viola and Jones, 2001) available in OpenCV (Bradski et al., 2006). In the following section, we present the methods used to extract facial features.

2.1 Feature Extraction Methods

Existing emotion recognition methods are mainly based on two types of features, namely geometric features and appearance features. For geometric features, we chose a distance-based method presented in (Abdat et al., 2011). Due to its face measure model, this method presents a good location of feature points independently of illumination changes and subjects changes. Moreover, it works in real time. For appearance features, we chose the Gabor method, a widely used method for texture extraction on different orientations and different scales.

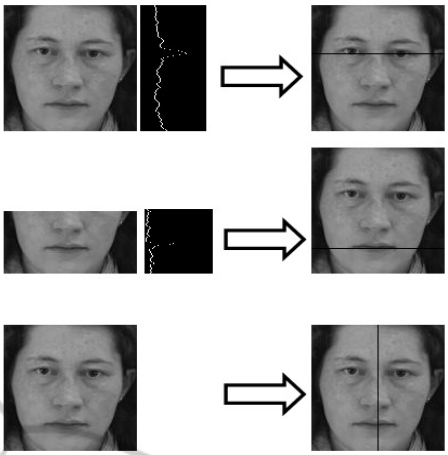

Figure 1: Techniques used to detect the three axis. The first row presents the horizontal projection of the horizontal gradient of the whole face, the second row presents the horizontal projection of the vertical gradient of the lower half of the face. The third row shows the location of the symmetric axis computed as the horizontal middle of the face.

2.1.1 Distance-based Method

Abdat et al (Abdat et al., 2011) developed a distance-based method. The facial expression is coded by distances variation linking the variation of the most relevant muscles to the human expressions. These distances are computed from a pair of dynamic and fixed points. The dynamic points are feature points that can move during the expression located on eyebrows, lips, eyelid and nose. The fixed points present stable points with respect to facial expression changes located on face edges, outer corners of the eyes and the nose root. The location of these points is based on the detection of the horizontal position of the eyes, the horizontal position of the mouth and the facial symmetric axis.

To improve the detection of these three axis, we changed some of the techniques used in (Abdat et al., 2011). For the detection of the eyes axis, the horizontal gradient projection is used (see the first row of figure 1). In our case, we use the Sobel mask to compute the horizontal gradient instead of columns difference. We also changed the mouth detection technique. Instead of using a HSV segmentation, we apply the horizontal projection of the vertical gradient. The second row of figure 1 illustrates the mouth axis detection, while the last row presents the symmetric axis detec-

tion which is computed as the horizontal middle of the face.

To ensure the position of feature points, the Shi&Tomasi method (Shi and Tomasi, 1994) is used in a neighbourhood of each point. In our case, we use a 8X8 block around each detected point. This method is available in the OpenCv library (Bradski et al., 2006).

The feature points are localised in the first frame of the image sequence which corresponds to the neutral face. Afterwards, these points are tracked using the Lucas-Kanade algorithm (Bouguet, 2000).

For each image, we obtain a distance feature vector composed of 21 distances.

2.1.2 Gabor Method

Gabor filter-based feature extraction has been successfully applied to fingerprint recognition (Lee and Wang, 1999), face recognition (Vinay and Shreyas, 2006) and facial feature point detection (Vukadinovic and Pantic, 2005). This is due to its similarity with the human visual system (Lee and Wang, 1999).

We applied the Gabor method to detect skin changes in each image. The faces were detected automatically and normalized to 80×60 sub-images based on the location of the eyes. The face is then filtered with a filter bank.

The entire filter bank can be generated by changing the orientation and the scale in the "mother" filter (1) (Kotsia et al., 2008a)

$$\psi_k(z) = \frac{||k||^2}{\sigma^2} \exp(-\frac{||k||^2||z||^2}{2\sigma^2})(\exp(ik^t z) - \exp(\frac{\sigma^2}{2})), \quad (1)$$

$z = (x, y)$ refers to the pixel and the wave vector $\vec{k}$ presents the vector of the plane wave restricted by the Gaussian envelope function, its characteristic: $k = [k_v \cos\phi_u, k_v \sin\phi_u]^t$ with $k_v = 2^{-\frac{v+2}{2}}\pi, \phi_u = \mu\frac{\pi}{8}$.

The parameter σ controls the width of the Gaussian $\frac{\sigma}{k}$, in our case $\sigma = 2\pi$. The subtraction in the second term of equation (1) makes the Gabor kernels *DC-free* to have quadrature pair (sine/cosine) (Movellan, 2005). Thus, the Gabor process becomes more similar to the human visual cortex. For our bank, we use three high frequencies for v=0,1,2 and four orientations $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$.

After the convolution of the face image with the Gabor bank, the face is again downsampled to 20×15. We obtain then a feature vector of 3600 descriptors ($20 \times 15 \times 12$).

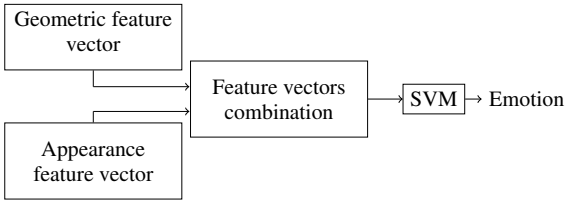

Figure 2: Early fusion scheme.

2.2 Geometric and Appearance Fusion Modalities

Geometric techniques and appearance approaches have their own strengths and limitations. The combination of both features may compensate the limitations of each method. The choice of the adequate fusion technique is also very important to enhance the emotion recognition system. Fusion can be done in the features space (early fusion) or in the decision space (late fusion). Early fusion combines weighted or equiprobable feature vectors in the same vector. Then, a classification method is applied. In contrast to early fusion, late fusion firstly applies a classification step to each feature vector independently and combines afterwards the obtained probabilities. In this paper, we studied various fusion methods.

2.2.1 Early Fusion Method

For each face image the geometric feature vector is extracted by the distance-based method ($X_G \in R^d$ with d=21 features) and the appearance feature vector is extracted by the Gabor method ($X_A \in R^{d1}$ with d1=3600 features). Both vectors are then normalized in $[0,1]$ using the Min_Max technique (Snelick et al., 2005). The minimum (*des_min*) and the maximum (*des_max*) of each descriptor are identified among all training vectors.

$$des\_norm = \frac{des - des\_min}{des\_max - des\_min} \quad (2)$$

A new feature vector is defined containing information from both geometric features and appearance features $X = [X_G, X_A]^T$. The feature vector X composed of 3621 descriptors is used as input to a linear Support Vector Machine (SVM). Figure 2 presents early fusion scheme.

2.2.2 Late Fusion Methods

Just like in the early fusion method the geometric and appearance feature vectors are first extracted for each face image. Then, a linear SVM classifier is applied to each feature vector to yield two posterior probability vectors $P(\omega_k|X_G)$ and $P(\omega_k|X_A)$ where ω_k is the

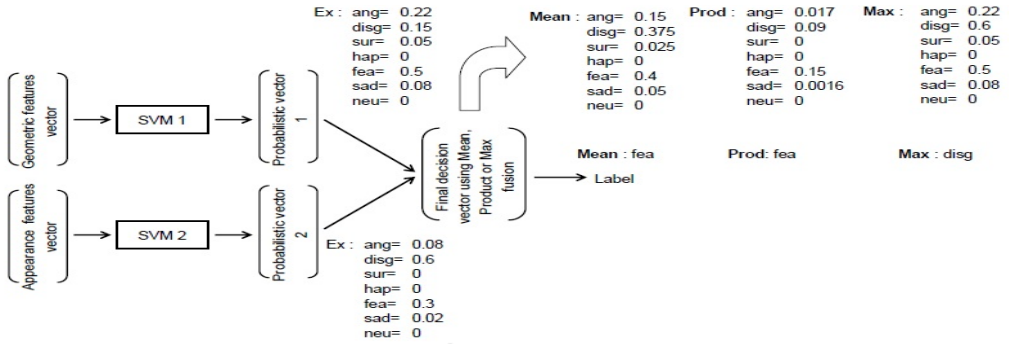

Figure 4: Examples of statistical fusion methods.

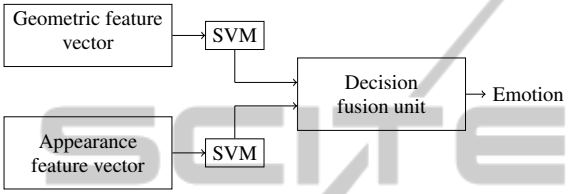

Figure 3: Late fusion scheme.

class of the emotion and $k \in \{1,...,n\}$, where n is the number of emotions. Those local decision vectors are then combined using a decision fusion step to obtain a final decision. This fusion scheme is illustrated in Figure 3.

We performed various modalities of decision fusion such as mean, product and maximum. A classification based-method (Atrey et al., 2010) has also been applied for this last decision fusion step. The next two sections are devoted to a more detailed presentation of the above mentioned decision fusion techniques.

2.2.3 Fusion by Statistical Rule

Various statistical rules exist for late fusion such as average, product, maximum, weighted majority voting, rank level (Mironica et al., 2013). We chose the most suitable techniques for our situation where a priori probabilities are not available.

Fusion by Average Rule

Under the equal prior assumption, the average of the obtained probability vectors is computed for each class. The maximum Mean is then selected as the final emotion as presented in the equation below. An example is shown in figure 4.

m represents the number of classification methods and $X_i \in \{X_G, X_A\}$ is a feature vector. These notations are used in the remainder of the paper.

$$Z \rightarrow \omega_k$$
$$\frac{1}{m}(\sum_{i=1}^{m} P(\omega_k|X_i)) = \max_k(\frac{1}{m}(\sum_{i=1}^{m} P(\omega_k|X_i)))|k = \{1,..,n\},$$

Fusion by Product Rule

We assume that the joint probabilities distribution measurements computed by SVM classifiers on each X_i are independent which means:

$$P(X_G, X_A|\omega_k) = P(X_G|\omega_k) \times P(X_A|\omega_k)$$

Under this assumption, the product rule is defined as:

$$Z \rightarrow \omega_k$$
$$if\ (\prod_{i=1}^{m} P(\omega_k|X_i)) = \max_k(\prod_{i=1}^{m} P(\omega_k|X_i)) \quad |k = \{1,..,n\},$$

Thus, the product of the obtained probabilities is computed for each class and the selected emotion is defined by the maximum product. An example illustrating this rule is presented in figure 4.

Fusion by Maximum Rule

The emotion is assigned to the maximal probability obtained in the decision vectors as explained below:

$$Z \rightarrow \omega_k$$
$$if\ \max_i(P(\omega_k|X_i)) = \max_k(\max_i(P(\omega_k|X_i)))$$
$$|k = \{1,..,n\}, i = \{1,.,m\},$$

An example is presented in figure 4.

2.2.4 Fusion by Classification Methods

Fusion by classification methods is mainly used in the domain of multimedia analysis (Snoek, 2005) (Niaz and Merialdo, 2013). A first learning step is applied to each feature vector to yield emotion scores, then these probabilistic scores are integrated to a second learning step to obtain the final emotion as illustrated in figure 5.

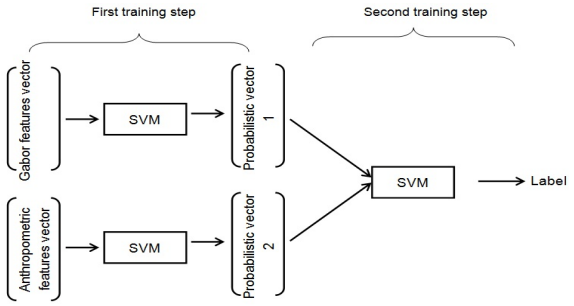

Figure 5: Classification-based fusion scheme.

The Support Vector Machine SVM classifier is applied in both learning steps, since it has many advantages namely low parameter number setting and fast training.

Two ways of training can be applied to the classification-based fusion methods. The first one uses just one training set which is applied in both training steps. The second one uses two different sets to train separately the first and the second training step. In this last case a large set of data must be available. In this paper, only the first way will be applied due to the reduced number of images available for each emotion in the considered databases.

3 DATABASES

Evaluation and comparison of these methods require the use of one or more databases. There are two types of databases: posed emotion ones and spontaneous ones. Posed emotion databases present forced emotions expressed by actors; while spontaneous databases present emotions stimulated by viewing videos. In the latter case, the emotions are often labelled according to the expected emotion; even if, in some cases, the expressed emotions are barely visible. In this paper, we chose an extended version of the widely used Cohn-Kanade database as forced expressions benchmark and the FEEDTUM database as spontaneous database.

The extended Cohn-Kanade database (CK+) contains facial expression videos from 123 subjects (an additional 26 subjects compared to Cohn-Kanade database) (Lucey et al., 2010). A total of 7 expressions are labeled including anger, contempt, disgust, fear, happy, sadness and surprise. The images presented in this database are digitalized into 640×490 pixels. The sequences vary from the neutral expression to the peak of the expression.

The FEEDTUM database is part of the European Union project FGNET (Face and Gesture Recognition Research Network) (Wallhoff, 2006). It contains face images and videos of 18 subjects performing the six basic emotions, stimulated by viewing videos. Each of them realizes the six emotions and the neutral expression three times. The images presented in this database are digitalized into 320×240 pixels. In total, it includes 399 sequences.

4 METHODS EVALUATION

The cross-validation method is a frequently used approach for performance evaluation. We use five fold cross-validation in which the data are randomly split into subsets of approximately equal size. Each set contains 20% of each emotion class. One set is chosen as a test set, while the remaining sets form the training set. After the classification step, the test set is integrated in the training set and a new test set is considered. This procedure is repeated five times. An average classification accuracy rate is then computed.

The cross-validation method is used to evaluate the hybrid methods in both CK+ and FEEDTUM databases. In the next section, a comparison between the performance of the fusion methods we chose and the performance of two hybrid methods presented in the literature is made.

4.1 Methods Comparison on the CK+ Database

A five fold cross-validation technique is applied to evaluate the recognition of the six emotions (anger (Ang), disgust (Dis), fear (Fea), happy (Hap), sadness (Sad) and surprise(Surp)) and the neutral expression (Neu) on the CK+ database.

4.1.1 Results Analysis

According to table 1, the distance-based method and the Gabor method have a similar mean recognition rate. The distance-based method achieves a recognition rate of 90.7%, while the recognition rate of the Gabor method reaches 90.4%. However, they do not misclassify the same emotion. In the case of the distance-based method the most misclassified emotion is sadness. In the case of appearance-based method the most misclassified situation is the neutral expression. The fusion of both features may correct these misclassifications.

The recognition rate of the early fusion method which combines the geometric and appearance features on the feature level, achieves 23% (see row 3 table 1). We notice that the early fusion method gives worse performance than the distance-based method

and the Gabor method when they are separately applied. This is due to the huge dimension of the Gabor vector compared to the geometric vector ($21 << 3600$). A feature selection method may be a good solution to improve the recognition rate of the early fusion method.

The late fusion methods based on statistical rules (average, product and max) are presented respectively in rows four, five and six of table 1. The recognition rates of these fusion methods are very similar. The three methods recognise very well happiness, sadness and surprise but classify worse fear. This emotion is jointly the third most misclassified emotion by the distance-based method and the second most misclassified emotion by the Gabor method. The other emotions have a good recognition rate because one of the two methods has a good recognition rate. Thus, the recognition rates of the statistical fusion methods which are closely linked to the response of the classifiers, are impacted. We conclude that the misclassification of the fear emotion by the individual classifiers affects the performance of the statistical fusion methods. Kuncheva (Kuncheva, 2002) reports that the difficult parts of the feature space are often the same for all classifiers. We remark that the statistical-based fusion methods improve the recognition rate of the emotions, more specifically the product-based rule fusion method. It enhances indeed all emotion recognition rates except anger which looses 4% compared to the other statistical-based fusion methods.

Classification-based fusion method is presented in the last row of table 1. The recognition rate of this method exceeds the recognition rate of the Gabor method and the distance-based method by approximately 3%. We notice also that it misclassified the neutral expression such as the Gabor method and unlike the distance-based method which achieves a rate of 100%. The classification-based fusion method has also a bad recognition rate for fear emotion. On the other hand, it has a good recognition rate for sadness and surprise. We can thus conclude that as the statistical-based fusion methods the classification-based method achieves good results when the Gabor method and distance-based method have good recognition rates. Similarly, the classification-based method misclassifies an emotion when both methods have bad recognition rates such as for the fear emotion. However, this method is also impacted when one of the classifiers has a bad recognition rate like for the neutral expression.

The comparison of the different fusion modalities shows that the late fusion methods prove to be a better choice than the early fusion in our task.

We notice also that the best recognition rates are given by the methods based on statistical rules for fusion. This is probably the reason why simple statistical rules continue to be mostly used for fusion approaches. An additional learning step does not have necessarily the best effect for emotion recognition application.

4.1.2 Comparison with Previous Work

A comparison of the proposed fusion method based on product rule and two methods of the literature that combine geometric and appearance features can also be done. We chose the Chen et al. (Chen et al., 2012) method which was initially intended to recognize seven emotions: happy, anger, fears, disgust, sadness, surprise and contempt using an early fusion technique to combine features and passing them to a SVM classifier. Kotsia et al. (Kotsia et al., 2008b) present also an early fusion method with the Median Radial Basis Function Neural Networks (MRBF NNs) to recognize six emotions (happy, anger, fears, disgust, sadness, surprise) and the neutral expression. They evaluate their method on the Cohn-Kanade database, first version of the CK+ database. The recognition rates of both methods are presented in table 2.

The proposed method exceeds recognition rate of 97% while Chen et al (Chen et al., 2012) method and Kotsia et al (Kotsia et al., 2008b) method only achieve respectively 95% and 92.3%. We also remark that the most misclassified emotion is fear for all methods. This is probably caused by the difficulty to simulate this emotion.

4.2 Spontaneous Expression Recognition on the FEEDTUM Database

According to psychologists, the difference between posed and spontaneous emotions is quite apparent. This difference is also highlighted in many computer vision application such as (Bartlett et al., 2006), (Zeng et al., 2009). To develop a real environment system, both emotion categories should be handled. This section is devoted to the evaluation of the previously considered methods in the recognition of spontaneous emotions. To this end, we use the FEEDTUM database which contains spontaneous and natural expressions. As expressions were captured under natural circumstances, head motion can be found in some sequences.

Table 3 presents the obtained recognition rates of the distance-based method, the Gabor method and the different fusion methods. We notice that the recognition rate of the Gabor method exceeds the recognition

Table 1: Fusion methods recognition rates computed by 5 fold cross-validation on the CK+ database.

	Methods	Recognition rates	Hap	Ang	Fea	Dis	Sad	Surp	Neu
Geometric	distance-based	90.7	96.0	89.5	87.0	85.3	83.0	93.7	100
Appearance	Gabor	90.4	97.7	88.0	83.7	96.0	93.3	98.0	75.7
Early fusion		23.0	14.0	50.6	24.8	31.5	4.0	36.6	0
Fusion based on statistical rules	Average	97.6	100	100	91.7	95.7	100	100	96.0
	Product	97.9	100	96.0	95.7	95.7	100	100	98.0
	Max	97.3	100	100	91.7	93.7	100	100	96.0
Fusion based on classification		93	95.7	92	83.7	98	100	100	81.7

Table 2: Performance of two emotion recognition systems from the literature which use appearance and geometric features.

Methods	Recognition rates	Hap	Ang	Fea	Dis	Sad	Surp	Neu
Chen et al (Chen et al., 2012)	95.0	97.5	92.5	90.0	96.0	93.5	96.5	_
Kotsia et al (Kotsia et al., 2008b)	92.3	97.5	93.6	84.3	89.5	94.3	95.6	91.3

Table 3: Fusion methods recognition rates computed by 5 fold cross-validation on the FEEDTUM database.

	Methods	Recognition rates	Hap	Ang	Fea	Dis	Sad	Surp	Neu
Geometric	distance-based	46.8	75.1	54.4	21.5	10.6	16.6	74.4	76.0
Appearance	Gabor	84.2	96.0	89.7	69.7	79.3	73.1	91.5	89.7
Early fusion		19.4	24.2	60.2	13.1	2.22	28.0	0	8.0
Fusion based on statistical rules	Average	83.3	100	85.5	65.5	75.5	70.8	95.5	89.7
	Product	83.9	100	81.5	72.0	77.5	72.8	93.3	89.7
	Max	84	98.0	89.5	65.3	75.5	71.1	97.7	89.7
Fusion based on classification		84	94	89.7	67.7	79.5	75.3	91.5	89.7

rate of the distance-based method from about 37%. For spontaneous expressions, the facial changes are often not clearly visible. Then, the resulting weak changes are hardly discernible in term of distances by the distance-based method. Besides, as mentioned above, during the expressions a head motion can also occur. The pretreatment done for the Gabor method consisting of scaling and normalising the face images based on the location of the two eyes, removes the head motion. On the other hand, the head motion affects the performance of the distance-based method.

We notice that the mean recognition rates of late fusion methods are very similar to the Gabor recognition rate. However, happiness and surprise are enhanced by the statistical-based fusion methods. This is due to the high recognition rates of both distance-based method and Gabor method in such emotions. We conclude that the recognition of positive spontaneous emotions which are more marked than the negative ones (fear, disgust...) are enhanced by statistical-based fusion methods. We remark also a slightly improvement in the recognition of the disgust and sadness by the classification-based method.

The comparison between the different fusion methods reveals that the decision level fusion methods are more reliable than the feature level fusion ones.

5 CONCLUSION

In this paper, various fusion methods are presented and developed to recognize posed and spontaneous facial emotions. Distance-based method and Gabor method extract respectively geometric features and appearance features. These features are combined in different levels (feature level and decision level). Fusion in the decision space proceeds either by statistical rules or by classification methods. Our test on the posed database (CK+ database) reveals that the statistical-based fusion methods are the most appropriate to recognize a greatly apparent expression. However on the spontaneous database (FEEDTUM database), the statistical-based methods enhance the recognition of the positive emotions. Besides, the classification-based method improves the recognition of sadness and disgust.

In future works, we intend to minimize the number of features used by the hybrid methods and enhance the recognition of the spontaneous emotions.

REFERENCES

Abdat, F., Maaoui, C., and Pruski, A. (2011). Human-computer interaction using emotion recognition from

facial expression. *5th UKSim European Symposium on Computer Modeling and Simulation (EMS)*, pages 196–201.

Anderson, K. and McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions Systems, Man, and Cybernetics*, 36(1):96–105.

Atrey, K., Anwar Hossain, M., El-Saddik, A., and Kankanhalli, S.-M. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia System*, pages 345–379.

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, pages 22–35.

Bartlett, M.-S., Gwen, L., Ian, F., and Javier, R.-M. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. *Computer Vision and Pattern Recognition Workshop*.

Bouguet, J. (2000). Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*.

Bradski, G., Darrell, T., Essa, I., Malik, J., Perona, P., Sclaroff, S., and Tomasi, C. (2006). *http ://sourceforge.net/projects/opencvlibrary/*.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.-M., Kazemzadeh, A., S., L., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. *6th International Conference on Multimodal Interfaces*, pages 205–211.

Chen, J., Chen, D., Gong, Y., Yu, M., Zhang, K., and Wang, L. (2012). Facial expression recognition using geometric and appearance features. *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, pages 29–33.

Fasel, I., Bartlett, M., and Movellan, J. (2002). A comparison of gabor filter methods for automatic detection of facial landmarks. *5th International Conference on automatic face and gesture recognition*, pages 345–350.

Gunes, H. and Piccardi, M. (2005). Affect recognition from face and body: Early fusion vs. late fusion. *IEEE International Conference on Systems, Man and Cybernetics*, 4:3437–3443.

Kotsia, I., Buciu, I., and Pitas, I. (2008a). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052–1067.

Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16:172–187.

Kotsia, I., Zafeiriou, S., and Pitas, I. (2008b). Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, pages 833–851.

Kuncheva, L. I. (2002). A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 281–286.

Lee, C.-J. and Wang, S.-D. (1999). Fingerprint feature extraction using Gabor filters. *Electronics Letters*, pages 288–290.

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., and Matthews (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion- specified expression. *IEEE Computer Vision and Pattern Recognition Workshops*, pages 94–101.

Mironica, I., Ionescu, B., P., K., and Lambert, P. (2013). An in-depth evaluation of multimodal video genre categorization. *11th International workshop on content-based multimedia indexing*, pages 11–16.

Movellan, J. (2005). Tutorial on gabor filters. *MPLab Tutorials, UCSD MPLab, Tech.*

Niaz, U. and Merialdo, B. (2013). Fusion methods for multi-modal indexing of web data. *14th International Workshop Image Analysis for Multimedia Interactive Services*, pages 1–4.

Shan, C., Gong, S., and Mcowan, P. W. (2009). Facial expression recognition based on Local Binary Patterns : A comprehensive study. *Image and Vision Computing*, 27:803–816.

Shi, J. and Tomasi, C. (1994). Good features to track. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, pages 593–600.

Snelick, R., Uludag, U., Mink, A., Indovina, M., and Jain, A. (2005). Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:450 –455.

Snoek, C. G. M. (2005). Early versus late fusion in semantic video analysis. *ACM Multimedia*, pages 399–402.

Vinay, K. and Shreyas, B. (2006). Face recognition using gabor wavelets. *4th Asilomar Conference on Signals, Systems and Computers*, pages 593–597.

Viola, P. and Jones, M. (2001). Robust real-time object detection. *In international journal of computer vision*.

Vukadinovic, D. and Pantic, M. (2005). Fully automatic facial feature point detection using gabor feature based boosted classifiers. *IEEE Conference of Systems, Man, and Cybernetics*, pages 1692–1698.

Wallhoff, F. (2006). Facial expressions and emotion database, http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html.

Wan, S. and Aggarwal, J. (2013). A scalable metric learning-based voting method for expression recognition. *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Zeng, Z., Pantic, M., Roisman, G.-I., and Huang, T.-S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, pages 39–58.

Zhang, L., Tjondronegoro, D., and Chandran, V. (2012). Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition. *IEEE International Conference on Multimedia and Expo*, pages 1027–1032.