TOWARD A MORE ROBUST FACIAL EXPRESSION RECOGNITION IN OCCLUDED IMAGES USING RANDOMLY SAMPLED GABOR BASED TEMPLATES

Ligang Zhang, Dian Tjondronegoro and Vinod Chandran

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

ABSTRACT

Occlusion is a big challenge for facial expression recognition (FER) in real-world situations. Previous FER efforts to address occlusion suffer from loss of appearance features and are largely limited to a few occlusion types and single testing strategy. This paper presents a robust approach for FER in occluded images and addresses these issues. A set of Gabor based templates is extracted from images in the gallery using a Monte Carlo algorithm. These templates are converted into distance features using template matching. The resulting feature vectors are robust to occlusion. Occluded eyes and mouth regions and randomly places occlusion patches are used for testing. Two testing strategies analyze the effects of these occlusions on the overall recognition performance as well as each facial expression. Experimental results on the Cohn-Kanade database confirm the high robustness of our approach and provide useful insights about the effects of occlusion on FER. Performance is also compared with previous approaches.

Index Terms— Facial expression recognition, face occlusion, Gabor, support vector machine

1. INTRODUCTION

Facial expression recognition (FER) plays an important role in many fields, including human computer interaction, surveillance, and multimedia content analysis. One of the major constraints in developing a robust FER system is facial occlusion, which often occurs in real situations. There are two types of facial occlusion: temporary and systematic [1]. Temporary occlusion can result from a part of the face obscured by a hand or an object or owing to head movement. Systematic occlusion results from wearing an item such as sunglasses, a scarf or mask.

It is still a challenge to overcome the influence of facial occlusion. Unlike other constraints such as pose variations, whose characteristics can be inferred beforehand, facial occlusion is particularly difficult to handle due to its "random" characteristic, which means that occlusions can occur at random positions and can be arbitrarily large in size. As a result, most of the existing approaches on FER just use non-occluded facial images taken under controlled laboratory conditions [2]. However, this is not the case in the real world.

Existing FER work in combating the challenge of occlusions reconstructs the occluded geometric features based on the configuration and visual properties of the face. These techniques include principal component analysis (PCA) [1], robust PCA [3], improved Kanade-Lucas tracker [4], [5] and transferable belief model [6]. However, they only adopt geometric features and lack the capacity to capture the regional appearance features (e.g. wrinkles), which also represent useful information for FER.

The effects of occlusion on recognition performance has been analyzed. These efforts either use both geometric and appearance features [7], [8], [9], or approach from the view of human perception [10], [11]. Although providing useful insights into the effects of occlusion on the performance (e.g. the most discriminative facial areas for different expressions), they have only used occlusions on the definite regions such as the mouth, eyes, nose, or left and right sides. To model the "random" characteristic of occlusions in the real world, it is necessary to test with occlusions at random locations and of different sizes. In addition, most training and testing data sets are matched and have been both occluded. It is of interest to test the recognition performance on occluded images when non-occluded images are used for training (mismatched conditions). This is particularly important for real applications in that it can eliminate the step of generating the occluded training images.

This paper proposes a robust FER approach to overcome these drawbacks using randomly sampled Gabor based templates. The templates are extracted by a Monte Carlo algorithm and serve as a pool of local features; therefore, only a part of them are influenced by occlusions. Template matching is used to find the most similar features located within a space around these templates. In this way, robust features are generated in the sense that occluded templates can be replaced by nearby non-occluded templates during matching. Five types of occlusion

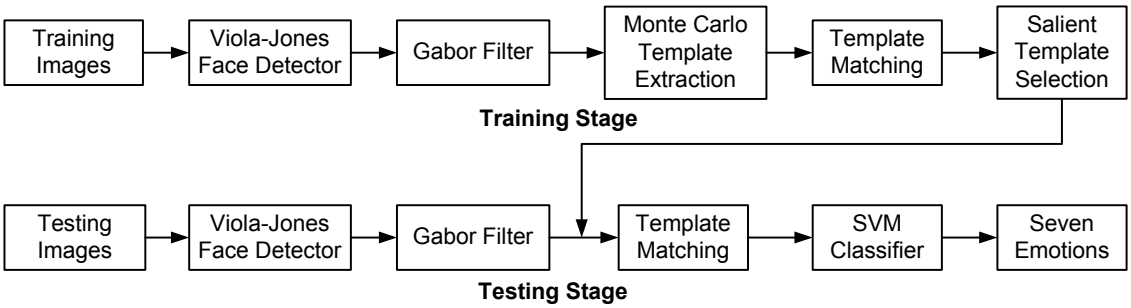


Fig. 1. Framework of the proposed approach.

(occluded eyes, occluded mouth, and three patch occlusions with random locations and different sizes) and two testing strategies (training and testing are both performed on occluded images i.e. matched; training is performed on non-occluded images and testing is done on occluded images i.e. mismatched) are adopted for the performance evaluation. Previous evaluations of FER have not considered random patches as in R_8, R_{16} and R_{24} tests here or the two testing strategies, to the best of our knowledge.

The rest of the paper is organized as follows. Section 2 presents the proposed approach. Section 3 describes the experiments. Finally, conclusions are drawn in Section 4.

2. PROPOSED APPROACH

As shown in Figure 1, the proposed approach is composed of a training stage and a testing stage. For an input training or testing image, the facial region is located by the widely used Viola-Jones face detector and scaled to 48×48 pixels. Occlusions are imitated by adding masks into the facial regions. Then Gabor images are obtained by convolving eight-scale (5:2:19 pixels), four-orientation (-45°, 90°, 45°, 0°) Gabor filters with the facial regions. The rest of the parameters of Gabor filters are set based on work [12]. During the training stage, a Monte Carlo algorithm is used to extract a set of templates, which are only partially influenced by occlusions. Template matching is then performed to find the most similar features located within a space around these templates, resulting in distance features robust against occlusions. Based on the distance features, a support vector machine (SVM) is used for selecting salient templates that are least influenced by occlusions. At the testing stage, the same template matching is performed on the salient templates to obtain the distance features in a testing image, which are fed into a SVM to recognize seven emotions: anger (AN), disgust (DI), fear (FE), happiness (HA), neutral (NE), sadness (SA) and surprise (SU).

2.1. Simulation of occlusions

No comprehensive public FER database has been reported with different types and sizes of facial occlusions. To simulate these occlusions, we add white masks to different positions of the face region, including (i) the two eyes, (ii) the mouth, (iii) randomly placed 8x8 patch R_8, (iv) R_{16} and (v) R_{24}. R_S means adding a mask with a size of S×S into the face region at a random location. This random patch technique has been used previously to simulate occlusions in face recognition [13]. The motivation for testing these occlusions stems from the real situations where glasses often occlude the two eyes, scarves and masks often occlude the mouth, and other unpredictable objects can occlude any part of the face randomly. In addition, eyes and mouth are widely regarded as the most important areas for facial expressions, therefore, occlusions of the eyes and mouth can be used to validate the robustness of the proposed approach. A set of occluded facial images is shown in Figure 2. It is worth mentioning that left/right face occlusions are not tested here as they have been shown to have little effect on the performance [8].

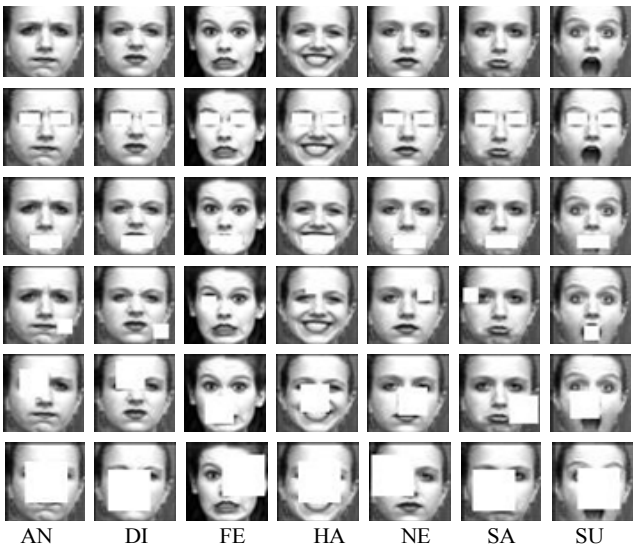


Fig. 2. Samples of occluded facial images. From top to bottom: no occlusion, occluded eyes, occluded mouth, R_8, R_{16} and R_{24}.

2.2. Template extraction

Template extraction adopts a Monte Carlo algorithm to collect a set of 3D templates from the occluded Gabor

images. These templates serve as a pool of local features and contain emotional information robust against occlusions. Figure 3 explains the algorithm. For each emotion and each template size, a number of (*Num)* templates are extracted. The final template set is obtained by combining the *Num* extracted templates of all emotions and all template sizes. To extract each of the *Num* templates, a five-step process is proposed: First, randomly select one image I from all training images with the emotion E_k. Second, randomly select one Gabor scale S_m from eight scales of the image I. Third, randomly select the X- and Y-axis positions from the scale S_m. Fourth, extract one template with a size of $P_j \times P_j \times O_{num}$ at the position of X and Y in the scale S_m. Finally, record the matching area and matching scale (details explained in Section 2.3).

Input: Template size P_j (j=1,…,4); emotion E_k (k=1,…,7); Gabor scale S_m (m=1,…,8); Gabor orientation number O_{num}.
Output: Templates, matching area and matching scale.

For each emotion E_k and each template size P_j
  For 1 to *Num* (*Num* = 1000)
    Randomly select one image I from training images of E_k;
    Randomly select one scale S_m from all scales in I;
    Randomly select X- and Y-axis positions in S_m;
    Extract templates with sizes of $P_j \times P_j \times O_{num}$;
    Record the matching area *Area*:
    Record the matching scale S_m;
  End
End

Return *templates, matching area* and *matching scale*.

Fig. 3. Algorithm for template extraction.

This algorithm is designed based on the "random" characteristic of facial occlusions and the fact that local features have an advantage over holistic features in solving occlusions [5]. In real situations, facial occlusions can randomly occur across the whole face region, but often influence only on a small part of the region. Therefore, it is expected that only a small proportion of the randomly extracted templates are affected by facial occlusions, while the majority of the templates obtained in the non-occluded regions remains uninfluenced. Similar algorithms have been used to extract a subset of features with a good generalization capacity of representing all features [12]. In this paper, four template sizes are used: 2×2×4, 4×4×4, 6×6×4, 8×8×4, and the *Num* is set to 1000. Given seven emotions, the final set contains 28,000 templates.

2.3. Template matching

Template matching aims to find the most similar features located within a space around the extracted templates. The matching helps to reduce the influence of occlusions because occluded templates can be replaced by nearby non-occluded templates. Thereby, robust features can be extracted. For each of the extracted templates, the matching produces a distance value by performing the following two steps, and the distances of all templates are concatenated to form the final feature set.

First, the matching area and matching scale are defined for each template to provide the matching space. The matching area has twice the width and height of each template but with the same orientation number and centre point. The matching scale is set as the same with the scale of P_a because the facial regions obtained by the Viola-Jones detector normally have the same scale.

Second, the distances between the template P_a and all templates within the matching space are calculated. Each calculation takes two templates as inputs and yields one value based on a distance metric. The minimum value in the matching space is the chosen feature for P_a. Two distance metrics are used: sparse L_1 (SL_1) and sparse L_2 (SL_2). The sparse metrics calculate the distances using the maximum value of all orientations in the template [14].

2.4. Salient template selection

SVM is used to select a subset of the most discriminative (termed "salient") templates. Based on the knowledge of the locations of occlusions learnt from the training images, we anticipate that the feature selection can eliminate those templates influenced the most by occlusions. SVM solves classification tasks by finding a hyperplane in a high dimensional space that can separate the negative examples from the positive examples with a maximum margin. The normal vectors to this hyperplane can be interpreted as feature weights that are updated every round to reflect the importance of each feature.

In this paper, a one-against-all SVM is trained for each emotion, and the average weights over the seven emotions are updated each training round based on the trained SVM. The features with low weights are dropped out (i.e. setting the weights to 0).

3. EXPERIMENTAL RESULTS

This section introduces the database used and analyzes the effects of different types of occlusions on the overall performance and each expression. Performance is also compared with previous approaches. The average correct recognition accuracy over 10 cross validations with a linear SVM classifier is used.

3.1. Database

The Cohn-Kanade (CK) database [15] is one of the most comprehensive benchmarks for facial expression tests. It includes 2105 digitized image sequences from 182 subjects. Six basic expressions were based on descriptions of prototypic emotions. Image sequences from neutral to

target display were digitized into 640×480 or 490 pixels.

In this paper, 1615 images that represent one of the seven emotions are selected from 92 subjects. For six basic emotions, the four peak images in each sequence are chosen. For the neutral case, the first images in four folds are chosen. The statistics of the chosen images are listed in Table 1. All images are classified into 10 similar sets and all images of one subject are included in the same set.

Table 1. Statistics of images chosen from CK database

	Total	AN	DI	FE	HA	NE	SA	SU
Image	1615	132	140	200	316	356	184	287
Subject	92	33	35	50	79	92	46	72

3.2. Effects of occlusion on overall performance

Figure 4(a) demonstrates the overall performance of the proposed approach under the five types of occlusions based on the matched train-test strategy. As can be seen, no occlusion (No) achieves the highest overall performance for all distances. The second best performance is for occluded eyes 95.1% accuracy using SL_1. The next ones are R_8, occluded mouth and R_{16}, whereas R_{24} ranks last. Compared to no occlusion, occlusion of the eyes leads to a small performance reduction (<1.7%), whereas occlusion of the mouth results in a big reduction of about 6%. When the size of random occlusions increases from 8×8 to 24×24, the overall performance has a significant reduction. However, the proposed approach still has 75% accuracy under R_{24}, for which one-half of the face is occluded. This demonstrates the high robustness of the proposed approach.

Figure 4(b) shows the overall performance of the proposed approach under five types of occlusion based on a mismatched train-test strategy. The proposed approach achieves a good performance under all occlusions except for occlusion of the mouth. Specifically, the best performers are still no occlusion, occluded eyes and R_8, while R_{16}, R_{24} and occluded mouth rank as the last three.

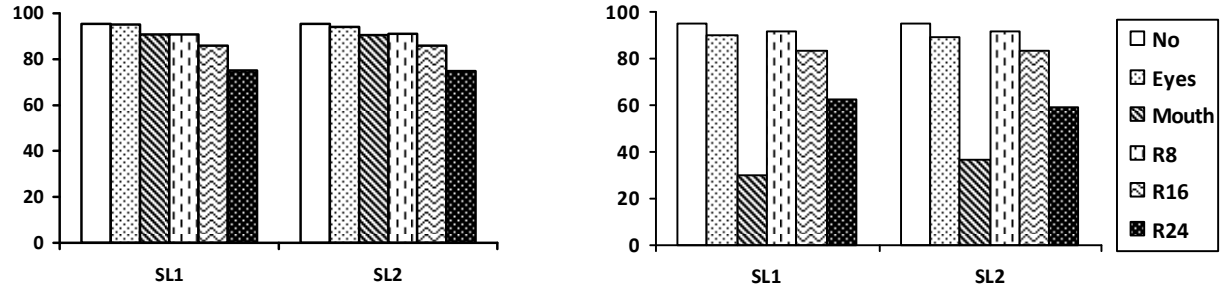
Compared with the results in Figure 4(a), most of the occlusions in 4(b) lead to performance reductions. Occlusion of the mouth leads to the biggest reduction for both distances (e.g. from 90.8% to 30.3% using SL_1). The one that follows in performance reduction is R_{24}, which suffers from a reduction from 75.0% to 62.5% when SL_1 is used. While occlusion of the eyes and R_{16} have small reductions of about 5% and 2% respectively. On the other hand, R_8 achieves little higher accuracy than that in Figure 4(a), which may be regarded as within a reasonable performance fluctuation.

As shown by the above results, for both the testing strategies, occlusion of the eyes and R_8 have little influence on the overall performance, while R_{24} and occlusion of the mouth exert a big influence. This implies that the *mouth region* and the *size of the occlusions* are two important factors determining the performance. The performance difference between the two testing strategies is owing to the fact that the SVM feature selector only selects features from non-occluded regions based on the knowledge learnt from training images during matched conditions, whereas there is no such knowledge when non-occluded images are used for training in the mismatched case.

3.3. Effects of occlusion on each facial expression

Since SL_1 and SL_2 produce similar performance for the two testing strategies and for most occlusions, the performance obtained by SL_1 alone is used for analyzing the effects of occlusions separately for each facial expression.

Figure 5(a) illustrates the effects of the five types of occlusion on each expression based on the matched train-test strategy. As can be seen, occlusion of the eyes has little influence on any expression. Furthermore, it slightly outperforms no occlusion (No) for anger, fear and sadness. This result is due to the performance fluctuations caused by the Monte Carlo algorithm of template extraction. Occluded mouth has a relatively significant effect on anger and sadness, while it has little effect on happiness, neutral and



(a) (b)

Fig. 4. Overall recognition performance (%) under five types of occlusion (a) when training and testing are both performed on occluded images (matched train-test strategy), and (b) when training is performed on non-occluded images and testing is done on occluded images (mismatched train-test strategy). "No" indicates no occlusion. SL1 and SL2 are two distance metrics. Note that occlusion of the mouth degrades performance in (b) significantly.

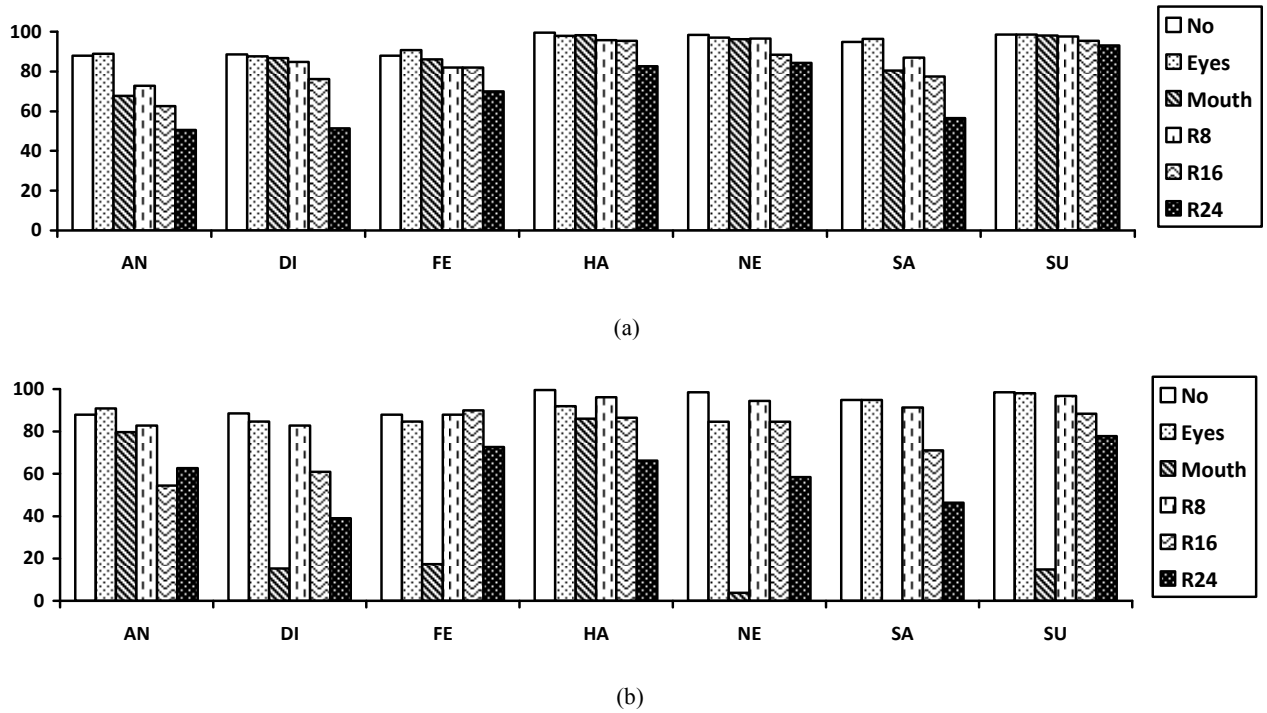


Fig. 5. Effects of five types of occlusion on the recognition performance (%) of each facial expression (a) when training and testing are both performed on occluded images (matched train-test strategy), and (b) when training is performed on non-occluded images and testing is done on occluded images (mismatched train-test strategy). Note that occlusion of the mouth degrades recognition performance of disgust, fear, neutral, sadness and surprise significantly but does not affect anger and happiness as much.

surprise. This result is reasonable for anger and sadness, but is contrary to the common understanding that the most distinguishing characteristic of surprise is an "open mouth". The reason may be that surprise on faces in the CK database is expressed not only by the mouth, but also by other regions. R_8, R_{16} and R_{24} pose a big influence on anger and sadness. In addition, R_8 and R_{16} also have a big effect on disgust. In contrast, the size of random occlusions has little effect on surprise.

Increasing size of random occlusions decreases the FER performance for all expressions, except fear, which is 82% accuracy under R_8 and R_{16}. All expressions except for surprise are most affected by R_{24}. In addition, anger and sadness are also strongly affected by R_{16} and occluded eyes. Neutral, surprise, disgust and happiness achieve similar performance under occlusions of the eyes, the mouth and R_8.

Figure 5(b) shows the effects on each facial expression based on the mismatched strategy. As shown in this figure, occlusion of the mouth demonstrates a significant impact on the performance of sadness, neutral, disgust, surprise, and fear, but a small effect on that of anger and happiness. R_{24} and R_{16} also have a big influence on all emotions, while R_8 influences little on all emotions. When the mouth is occluded, all sad images are misclassified, and this agrees with the result in [4] that only 20% accuracy is obtained for sadness. Moreover, neutral, disgust and surprise obtain around 3.8%, 15%, and 15% accuracy respectively, while happiness and anger obtain 86.1% and 79.8% accuracy respectively. It is interesting to note that higher accuracy is obtained for anger under R_{24} than under R_{16} (62.6% versus 54.6%).

Compared with the results in Figure 5(a), occlusion of the mouth, R_{24} and R_{16} suffer big performance reductions. On the other hand, occlusion of the eyes and R_8 perform similarly for the two testing strategies. For matched train-test conditions, all emotions are mostly affected by R_{24}, whereas for mismatched conditions all emotions except for anger and happiness are mostly affected by occluded mouth. Similar to the results presented in [7], [8], occlusion of the mouth is observed to result in a bigger performance reduction than occlusion of the eyes for most expressions.

3.4. Comparison with previous approaches

We also evaluate the proposed approach by conducting a performance comparison with previous approaches, which adopted Gabor features and tested seven expressions on the CK database. To make a fair comparison, we only compare the performance under occlusions of the eyes and the mouth based on the matched strategy. It should be noted that no research has been reported that tests FER performance based on the mismatched strategy. As can be seen from the results in Table 2, the proposed approach outperforms two benchmarked approaches. A shortcoming of the

benchmarked approaches is that they extract a Gabor feature vector from facial regions without eliminating features influenced by occlusions. On the other hand, the proposed approach randomly extracts a set of templates from facial regions to represent local features, performs a template matching to find robust features within a space, and utilizes the feature selection to remove the templates influenced most by occlusions.

Table 2. Performance (%) comparison with previous approaches

	Approaches	Eyes	Mouth
Proposed	Template Gabor + SVM	95.13	90.76
[7]	Gabor filter (Lfr) + MCC	92.3	87.2
	Gabor filter (Lfr) + CSM	91.5	86.4
[8]	Shape + SVM	88.4	86.7
	Gabor filter	86.8	84.4
	DNMF	84.2	82.9

4. CONCLUSION

This paper explores facial expression recognition in occluded images. We use a Monte Carlo algorithm to select Gabor templates from the gallery images and template matching over a search area to generate features robust against occlusions. 75% recognition accuracy is achieved for train-test matched conditions even when a half of the face is occluded. For train-test mismatch, the accuracy is 62.5%. Occluded eyes and random 8x8 (or R_8) patches have little effect, while occluded mouth and R_{24} lead to big performance reductions. For train-test matched conditions all emotions are mostly affected by R_{24}. For mismatched train-test conditions, occlusion of the mouth has a significant effect on all emotions except anger and happiness. The results also indicates that better FER performance can be achieved with occluded facial images if the training set also includes images with the same type of occlusion.

5. REFERENCES

[1] H. Towner and M. Slater, "Reconstruction and Recognition of Occluded Facial Expressions Using PCA," in *Affective Computing and Intelligent Interaction*, 2007, pp. 36-47.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 39-58, 2009.

[3] M. Xia, X. YuLi, L. Zheng, H. Kang, and L. ShanWei, "Robust facial expression recognition based on RPCA and AdaBoost," in *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, 2009, pp. 113-116.

[4] F. Bourel, Chibelushi, C.C., Low, A.A., "Recognition of facial expressions in the presence of occlusion," *In: 12th British Machine Vision Conference,* vol. 1, pp. 213-222, 2001.

[5] F. Bourel, C. C. Chibelushi, and A. A. Low, "Robust facial expression recognition using a state-based model of spatially-localised facial dynamics," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, 2002, pp. 106-111.

[6] Z. Hammal, M. Arguin, and F. Gosselin, "Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions," *Journal of Vision,* vol. 9, pp. 1-19, 2009.

[7] I. Buciu, I. Kotsia, and I. Pitas, "Facial expression analysis under partial occlusion," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. 453-456 Vol. 5.

[8] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing,* vol. 26, pp. 1052-1067, 2008.

[9] Z. Yongmian and J. Qiang, "Active and dynamic information fusion for facial expression understanding from image sequences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 699-714, 2005.

[10] J. N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology,* vol. 37, pp. 2049-2058, 1979.

[11] M. Nusseck, Cunningham, D. W., Wallraven, C., & Bülthoff, H. H., "The contribution of different facial regions to the recognition of conversational expressions," *Journal of Vision,* vol. 8(8):1, pp. 1-23, 2008.

[12] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 29, pp. 411-426, 2007.

[13] K. Jongsun, C. Jongmoo, Y. Juneho, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 1977-1981, 2005.

[14] J. Mutch and D. G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 11-18.

[15] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000, pp. 46-53.