Towards a dynamic expression recognition system under facial occlusion

Xiaohua Huang\textsuperscript{a,b,*}, Guoying Zhao\textsuperscript{a}, Wenming Zheng\textsuperscript{b}, Matti Pietikäinen\textsuperscript{a}

\textsuperscript{a}Center for Machine Vision Research, Department of Computer Science and Engineering, University of Oulu, Oulu, 90014, Finland
\textsuperscript{b}Key Laboratory of Child Development and Learning Science (Ministry of Education), Southeast University, Nanjing, JiangSu 210096, China

ABSTRACT

Facial occlusion is a challenging research topic in facial expression recognition (FER). This has resulted in the need to develop some interesting facial representations and occlusion detection methods in order to extend the FER to uncontrolled environments. It should be noted that most of the previous work focuses on these two issues separately, and on static images. We are thus motivated to propose a complete system consisting of facial representations, occlusion detection, and multiple feature fusion in video sequences. For achieving a robust facial representation, we propose an approach deriving six feature vectors from eyes, nose and mouth components to form a facial representation. These features with temporal cues are generated by the dynamic texture and structural shape feature descriptors. On the other hand, occlusion detection is still mainly realized by the traditional classifiers or model comparison. Recently, sparse representation has been proposed as an efficient method against occlusion, while it is correlated with facial identity in FER, unless using an appropriate facial representation. Thus, we present an evaluation demonstrating that the proposed facial representation is independent of facial identity. Inspired by Mercier et al. (2007), we then exploit the use of the sparse representation and residual statistics to occlusion detection of the image sequences. As concatenating six feature vectors into one causes the curse of dimensionality, we propose multiple feature fusion consisting of fusion module and weight learning. Experimental results on the Extended Cohn–Kanade database and its simulated database demonstrate that our framework outperforms the state-of-the-art methods for FER in normal videos, and especially, in partial occlusion videos.

1. Introduction

The goal of automatic facial expression analysis is to determine the emotional state, e.g. happiness, sadness, surprise, neutral, anger, fear, and disgust of human beings based on the facial images, regardless of facial identity. Thus far, most researchers have conducted experiments under controlled environments, such as un-occluded face (Fasel and Luettin, 2003; Zeng et al., 2009). Unfortunately, in many scenarios, the human’s face may be partially occluded. For example, a sunglass or a virtual reality mask occludes eyes, while a scarf or medical mask occludes the mouth and nose. Facial occlusion leads to facial representation or occlusion detection methods in combating the challenge of occlusions.

As known, facial-representation-based approaches are proposed in the literature to describe the appearance and/or geometric features for analyzing facial expressions. During the past decades, some popular representation methods have been used on facial occlusion in facial expression recognition (FER), such as Gabor or eigenfaces. Buciu et al. (2005) proposed the use of Gabor feature and fusion of classifier output on facial occlusion. Recently, Zhang et al. (2011) proposed the use of the template matching based on Gabor for recognizing the expressions with occlusion. Towner and Slater (2007) used eigenface on facial points for reconstructing incomplete facial expressions and then recognizing them. All these methods encoded the entire facial image as one feature (Towner and Slater, 2007; Buciu et al., 2005; Zhang et al., 2011). However, not all facial parts are of the same importance to facial expressions. It thus leads to the kinds of methods that decompose the facial image into sparse facial parts (Kotsia et al., 2008; Zhi et al., 2011). Algorithms like that can achieve a sparse decomposition through the salient facial features.

Recently, except facial-representation-based approaches, an alternative named as occlusion detection, has been proposed to automatically detect facial occlusion. In this method, occlusion detection is usually realized through three procedures. It firstly models facial representation on the un-occluded faces, and then detects occlusion that is not explained by normal facial representation, finally, it only processes the un-occluded regions of face (Min et al., 2011). For example, robust principal component analysis and

* Corresponding author at: Key Laboratory of Child Development and Learning Science (Ministry of Education), Southeast University, Nanjing, JiangSu 210096, China. Tel.: +86 358 85337565; fax: +86 358 85332612.

E-mail addresses: xiaohua.huang@seu.edu.cn (X. Huang), guoying.zhao@seu.edu.cn (G. Zhao), wenming.zheng@seu.edu.cn (W. Zheng), mki@ee.oulu.fi (M. Pietikäinen).
Saliency detection were used to detect facial occlusions (Mao et al., 2009). In Mercier et al. (2007), it was proposed to construct the model of the un-occluded face by a fitting algorithm, and then detect occlusion by means of residual image statistics. It is observed that occlusion detection has good performance in dealing with occlusion.

Although the above mentioned approaches tried to handle the occlusion of FER, it is well known that most works described in the literature are done on the static images under occlusion (Mercier et al., 2007; Towner and Slater, 2007; Buciu et al., 2005; Kotsia et al., 2008; Zhang et al., 2011; Zhi et al., 2011; Mao et al., 2009). In Fasel and Luettin (2003) and Zeng et al. (2009), the surveys have shown that dynamic features from the image sequences can provide much more information than the static images. Recently, some methods for image sequences have been proposed to solve the occlusion. In Bourel et al. (2000), Bourel et al. used six local facial regions and a rank-weighted k-nearest-neighbor classifier against partial occlusion in videos. In Fanelli et al. (2010), a Hough Forest-based method was proposed for facial expression sequences. It means that even though the facial sequence is partially occluded, the un-occluded regions in the video can still provide the motion information to FER. Compared with works on static images, there are few works on the video sequence under occlusion by now. Furthermore, some methods, such as Gabor and eigenfaces, are difficult to extend to the video sequences. Even though a model of un-occluded image sequences is constructed, the model is so complicated that it cannot handle occluded regions easily.

In facial representation, it is important to explore a robust feature to occlusion. Instead of presenting their robustness, most of these methods intend to investigate the effects of occlusion on facial expression. In other words, few studies try to construct a robust feature to describe the facial expression for the purpose of combating with facial occlusion. The aim of facial-representation-based approach in this paper is thus to compress the noise mostly. Inspired by Kotsia et al. (2008) and Zhi et al. (2011), we propose the use of a component-based approach to represent facial expressions. Its advantage is that some facial regions are considered separately so that the occluded region cannot produce too much noise to an un-occluded region. Moreover, the advantage of occlusion detection (Mercier et al., 2007; Mao et al., 2009) is that it can detect occlusion and then discard occlusion. Motivated by the idea of occlusion detection, we aim to detect the occlusion of each component in order to further exploit the performance of un-occluded facial parts. It is found that most of researches concerning occlusion consider the facial representation or occlusion detection solely for FER. In other words, few studies are conducted on combing facial representation and occlusion detection. We thus study a complete framework of dynamic expression recognition under facial occlusion, as shown in Fig. 1(a). It consists of three essential parts: (1) a component-based spatiotemporal feature descriptor,
especially robust to occlusion, (2) the occlusion detection based on sparse representation theory, and (3) multiple feature fusion.

To our knowledge, this is the first time to propose a complete framework for dealing with facial occlusion in video sequences. Compared with traditional methods, our framework has the following advantages: (1) Facial representation and occlusion detection are two separate issues in previous studies, but we take both of these into account. (2) The traditional methods are based on static images, while our framework is based on video sequences.

The proposed framework was evaluated on the Extended Cohn–Kanade facial expression database (CK+) (Lucey et al., 2010). The original Cohn–Kanade (Kanade et al., 2000) includes 486 FACS coded sequences from 97 subjects for six basic expressions: happiness (Ha), sadness (Sa), surprise (Su), anger (An), disgust (Di) and fear (Fe). For CK+ distribution, it has been further augmented to include 593 sequences from 123 subjects for seven expressions (107 sequences, 26 subjects and contempt expression (Co)), which makes it more challenging than the original database.

In our experiments, 325 sequences from 118 subjects were selected from the database for seven basic expression recognitions. The Leave-One-Subject-Out method was used in the whole scenario.

The rest of the paper is organized as follows: the component-based facial expression representation is presented in Section 2. Sparse-representation-based occlusion detection is presented in Section 3. In Section 4, multiple feature fusion and weight learning are described. Experiments and discussion can be found in Section 5. In Section 6, the conclusion are drawn. Preliminary results of part of this work have been published in Huang et al. (2011).

2. Component-based multiple facial representation

In Sun and Yin (2009) and Zhang and Ji (2005), the extension of component-based approach to facial sequence by the time development demonstrates a good performance in FER. Furthermore, the combination of multiple geometric and appearance features can describe the face in a better way (Fasel and Luettin, 2003; Lucey et al., 2010; Li et al., 2009). In Lucey et al. (2010), both similarity-normalized shape and canonical appearance were derived from the active appearance model to interpret the face image. As different regions of a face perform correlatively with certain relations during an expression action, we thus proposed the component-based multiple feature vectors for facial representation, which consists of the appearance and shape representations. The procedure of the component-based facial expression representation is shown in Fig. 1(b).

2.1. Component-based Multiple facial representations

Different from global method or block-based method, on the facial image, we consider the component-based approach. In this method, for facial representation, three facial components (mouth, nose and eyes) are considered (Huang et al., 2011) according to the facial configuration, and then feature vectors are extracted from three facial components.

Recently, the local binary pattern (LBP) operator, which is a gray-scale invariant texture primitive statistic, has shown excellent performance in the classification of various kinds of textures (Ojala et al., 2002). The dynamic version of LBP (Zhao and Pietikäinen, 2007) effectively describes the appearance (XY plane), horizontal motion (XT plane) and vertical motion (YT plane) from the image sequence. Those inspired us to utilize LBP in facial expression. In Fig. 1(b), we consider LBP on three facial components, and then the uniform patterns (Ojala et al., 2002) that at most two bitwise transitions from 0 to 1 or vice versa occur when the binary string is considered circular is utilized in each plane for dimensionality reduction. In our experiments, the radii in axes X, Y, and T are set as three; the number of local neighboring points around the central pixel for all three planes is set as eight.

For representing facial shape, Edge map (EdgeMap) (see Gizatdinova and Surakka, 2006 for details) were utilized as describing structural features, in which histogram statistically similar to LBP. Here, we briefly review the theory of EdgeMap. Given a smoothed image with a set of 10 kernels, the whole set of 10 kernels results from the differences between two oriented Gaussians with shifted kernel:

\[ G_l = \sum_{x,y} (|G_l - G_0|)^r \cdot h |G_l - G_0|^s, \]

where

\[ G_0 = \frac{1}{2\pi\sigma^2} \exp \left( -\frac{(u - \sigma \cos \theta_i)^2 + (v - \sigma \sin \theta_i)^2}{2\sigma^2} \right), \]

\[ G_l = \frac{1}{2\pi\sigma^2} \exp \left( -\frac{(u + \sigma \cos \theta_i)^2 + (v + \sigma \sin \theta_j)^2}{2\sigma^2} \right), \]

and \( \sigma \) is a root mean square deviation of the Gaussian distribution, \( \theta_i \) is the angle of the Gaussian rotation, \( \theta_j \) is estimated by the orientation of a local edge \( \theta_{ij} = \sum_{u,v} g_{l+1} \cdot g_{l+1}, \) where \( g_{l+1} \) denotes the gray level of the image at the pixel \( (i,j); l = 0, \ldots , W - 1, j = 0, \ldots , H - 1; W \) and \( H \) are the width and height of the image, respectively. After getting the edge orientations for an image, a histogram is created to collect up the occurrences of different orientations. In our approach, we extend EdgeMap to describe the shape representation for three components in an image sequence. As shown in Fig. 1(b), if we concatenate the histograms from all the frames, the videos of different length would have different sizes due to their arbitrary length. To solve this problem, we adopt the mean method (Wu et al., 2010) to encode the structural feature of one video sequence. Thus, the structural feature for one video sequence is represented by one histogram.

2.2. Discussion

Studies in Zhao et al. (2010) proposed the use of EdgeMap to describe the structural features and LBP to describe the motion features. Inspired by it, we also utilize EdgeMap to describe the structural features, while the difference includes (1) that we extract the structural features by EdgeMap and the dynamic appearance via LBP simultaneously in the XY plane, (2) that we extend EdgeMap into the sequence through a statistical method, and (3) that we use a trick of multiple feature fusion (see Section 4 for details) to the sequence for the purpose of reducing the mutual effect of each region.

3. Occlusion detection

Occlusion detection is used to automatically detect the occlusion. Some works (Mercier et al., 2007; Mao et al., 2009) have shown that occlusion detection has good performance in dealing with occlusion. Its basic idea is to compare the testing sample to the model on the un-occluded face. Recently, sparse representation has been a popular way to handle the occlusion in face recognition.
(Wright et al., 2009). Motivated by it, we also propose occlusion detection, which is based on sparse representation for FER.

3.1. Sparse representation

Given training samples of the i-th object class, each of the training samples is denoted as \( x_i \in \mathbb{R}^D \), where \( D \) is the dimensionality, \( j \) represents the index of a training sample of the i-th object class. Let us define \( X = [x_1, \ldots, x_C] \), where \( C \) represents the number of the object class. Given any test sample \( x_{test} \in \mathbb{R}^D \) from the same class can be linearly spanned by the training samples of the i-th object class,

\[
x_{test} = \omega_i^0 x_1^0 + \omega_i^1 x_1^1 + \cdots + \omega_i^n x_n^i = x_i \omega_i,
\]

where \( \omega_i^0 \) is a reconstruction coefficient of the i-th object class and the training sample, and \( \omega_i = [\omega_i^0, \ldots, \omega_i^n] \).

As can be seen, the formulation (2) pursues the minimization of the reconstruction of the training samples. This problem generates some kinds of facial representations. The most famous representation is the class specific PCA (Belhumeur et al., 1997) based on the \( l_2 \)-optimization problem. It is done by minimizing the objective function of the reconstruction error. Recently, sparse representation (Wright et al., 2009) has been proposed to find an object representation through a sparse combination of the samples in the gallery. Comparing with PCA, sparse representation is achieved by optimizing an objective function that consists of the reconstruction error and the sparsity of the representation. In general, an \( l_1 \)-optimization problem can reach an achievement of the sparsity of the representation. Thus, the optimization problem of finding the sparse vector \( \omega \) is as follows:

\[
\hat{\omega} = \arg \min \| \omega \|_1 \quad \text{subject to} \quad \| X \omega - x_{test} \|_2^2 \leq \epsilon,
\]

where \( \epsilon \) is the threshold, and vector \( \omega \) depicts the contribution of the training images.

3.2. Extension to facial expression recognition

In sparse representation (Wright et al., 2009) for face recognition, the sparse vector of weight acquired high responses to images of the same identity class and low ones to the other images. It motivates us to extend sparse representation to the reconstruction of facial sequence. Recently, the research in Zafeiriou and Petrou (2010) has shown that the sparse representation for FER is hardly achieved from original images, and sparse coefficient vectors are much more related to the facial identity than to facial expression. Thus, we analyze and discuss the features for facial sequences regarding whether they satisfy with the sparse condition and irrelevance to facial identity.

In Section 2, we assume that there are \( C \) expression classes and \( N \) training sequences sorted by the order of expression classes, and each image is divided into \( M \) components, therefore, dynamic appearance and statistical structure representations for training sequences are represented as \( \{L_{mj}\}_{m=1}^{M} \) and \( \{E_{mj}\}_{m=1}^{M} \) respectively. For a given facial sequence with unknown facial expressions, its dynamic appearance features and statistical structure features are extracted via the component-based approach, and are also denoted as \( \{L_{mj}\}_{m=1}^{M} \) and \( \{E_{mj}\}_{m=1}^{M} \). In order to increase the robustness of occlusion detection, we concatenate two features \( L_{mj} \) and \( E_{mj} \) for the \( m \)-th component in the \( j \)-th sequence into a combined feature \( F_{mj} \). The same procedure was done in the test sequence, which is denoted as \( F_{mj}^{test} \). After that, the problem (3) can be solved by Donoho et al. (2005).

The results of sparse representation of three components in the person-independent and person-dependent experiments are shown in Fig. 2. As can be seen in the figures on the first two rows (the person-independent experiment), we achieved high value responses for the same expression from the training samples at most of the components. In the figures on the last two rows (the person-dependent experiment), most of the components can achieve a high value response for the same expression from the training samples, with the exception of the feature vector of the nose in the surprise response to the same person with different expressions. We can thus consider that (1) the decomposition is sparse and produces high value responses for the same expression, and (2) the decomposition of features is person-independent.

3.3. Toward occlusion detection

The aim of occlusion detection is to determine the region whether or not occluded. With the evaluation and conclusion in the previous section, we thus exploit sparse representation to develop a system to detect the occlusion for facial expressions. According to the problem (3), for a given partially occluded sequence \( F_{mj}^{test} \), the coefficient vectors \( [\omega_0^1, \ldots, \omega_0^n] \) are ordered and obtained by the index of training samples. Using only the coefficients associated with the i-th expression class, we approximate \( F_{mj}^{test} \) as \( F_{mj}^{test} = F_{mj} \omega_i \). In our paper, the residual between \( F_{mj}^{test} \) and \( F_{mj} \) in the i-th expression class is defined as \( r_i(F_{mj}^{test}) = \| F_{mj}^{test} - F_{mj} \|_2 \), and the residuals associated with all expression classes of the test sample are represented by \( \{r_i\}_{i=1}^{C} \). Furthermore, the minimum residual \( r_{min} \) is chosen, i.e. \( \min \{r_i\}_{i=1}^{C} \).

It is noted that the test sample can be represented by the samples associated with the i-th expression class. However, the test sample with occlusion causes larger residual than the un-occluded sample. For classifying occlusion, the threshold machine is defined as:

\[
h(x_{test}) = \begin{cases} 0, & r_{min} \geq \gamma, \\ 1, & r_{min} < \gamma, \end{cases}
\]

where \( \gamma \) is the threshold of residual, ‘0’ and ‘1’ represent the occluded and un-occluded status, respectively. In Wright et al. (2009), the recognition was improved by block partitioning. Since one component is not absolutely occluded, the block-partition-based method is introduced to detect each block of the component. In the following experiments, the threshold is set as 0.2.

3.4. Procedure of occlusion detection

The procedure of occlusion detection for a facial sequence is divided into two stages, (1) the procedure of codebook generation and (2) the feature regroup. Fig. 3 shows the procedure of two stages.

In the first stage, we concatenate two feature vectors in the same volume (shown in Fig. 3(a.1)) into one feature (shown in Fig. 3(a.2)) in order to strengthen the robustness of occlusion detection. For each volume, the reconstruction residual is computed (shown in Fig. 3(a.3) and (a.4)). Through the threshold, the codebook for the component (shown in Fig. 3(a.6)) is generated for the next stage. And this codebook shows whether the regions are occluded or not. For example, the residual values of three regions associated with all expression classes are shown in Fig. 3(a.4). It is observed that the 1st volume and 30-th volume have smaller minimized residual values, although the 30-th volume is partially occluded by eyeglasses. However, the 71-st volume is occluded by a major part of eyeglass, and its minimum value is 2.1649. Compared with the 1st and 30-th volumes, all residuals of the 71-st volume are greater. It means that the definition of residual is meaningful to occlusion detection. Through the threshold, the codebook for these three volumes is 1, 1, and 0.
In the second stage, it is important to remove the occlusion parts in the sequence. In the feature representation, there are dynamic appearance features and statistical structural features for each component, as shown in Fig. 3(b.1). The codebook (shown in Fig. 3(b.2)) is attributed appropriately to each of two features, so that the occluded regions in each component are discarded. Finally, each regrouped feature of the two features is shown in Fig. 3(b.3).

4. Multiple feature fusion

How to fuse multiple feature descriptors is an important issue. Generally, two or more features are usually concatenated into a single feature. Here, we call this case as the feature-integration approach. It is noted that it will cause the curse of dimensionality to the classifier. Alternatively, we exploit a fusion method, i.e. the
mean rule (see Kittler et al., 1998, for details), for integrating multiple feature vectors. Here, we call it as the feature-fusion approach. The approach in our paper can achieve three goals which are: (1) simplify the computation of classification, (2) allow each feature to play its key role in facial expression and (3) reduce the mutual effect of each region.

Assume that all features vectors \( \{ \Omega_m \}_{m=1}^M \), where \( M = 2 \times M \), are generally statistically independent, and the prior probability of occurrence for the \( c \)-th class model is under the assumption of equal prior. The fusion rule of multiple feature sets is described as: Assign \( x_{\text{test}} \rightarrow \mu \) if
\[
P(c = \mu | x_{\text{test}}, \Omega_1, \ldots, \Omega_M) = \max_{c=1, \ldots, C} \left\{ P(c(\Omega_1) \beta_{1,c}, \ldots, P(c(\Omega_M) \beta_{M,c}) \right\},
\]
where \( x_{\text{test}} \) and \( \mu \) represent the testing sample and its class, respectively. \( \beta_{m,c} \) represents the weights of the \( m \)-th feature vector to the \( c \)-th expression. In multiple feature fusion, LIBSVM (Chang and Lin, xxxx) is used to generate voting probabilities. It is achieved by the following: (1) We divide multi-class classification into multiple binary classifications. (2) For the \( m \)-th feature vector, SVMs generate the voting number \( V_{c,m} \) for the \( c \)-th expression. (3) For all expressions, the total voting number is defined as the sum of \( V_{c,m} \), \( c = 1, \ldots, C \), denoted as \( V_{\text{m}} \). Thus, the probability of the \( m \)-th feature vector with the \( c \)-th expression is defined by \( V_{c,m} / V_{\text{m}} \). It is denoted as \( P(c_{m} \beta_{m}) \). The binary SVM model for each feature vector is trained by LIBSVM in the training step.

Recently, the MKL has been introduced to combine heterogeneous sources of information for decision fusion in computer vision (Dileep and Sekhar, 2009; Goene and Alpaydin, 2009). Thus, we exploit the use of the MKL to learn optimal weights \( \beta_{m,c} \) for multiple feature vectors.

Given multiple feature vectors \( \{ \Omega_m \}_{m=1}^M \), each of them has \( N \) samples \( \{ x_{mi} \}_{i=1}^N \), and the corresponding class label of \( x_{mi} \) is \( y_i \), where \( y_i \in \{+1, -1\} \), one can calculate multiple basis kernels for each feature vector. Hence, the kernel of the MKL is computed as a convex combination of the basis kernels
\[
k_{ij} = \sum_{m=1}^M \beta_{m} k(x_{mi}, x_{mj}),
\]
and restricted by \( \sum_{m=1}^M \beta_m = 1 \), where \( \beta_m \) is the weight of the \( m \)-th feature vector, \( k(x_{mi}, x_{mj}) \) is the kernel of \( x_{mi} \) and \( x_{mj} \). The formulation (6) is to show the purpose of the MKL that combines the multiple feature vectors into a single feature vector by assigning different weights. It thus is defined based on all components. Here, we use a linear kernel function to compute the basis kernel. And then we normalize all kernel matrices of feature sets to unit trace. To make sure each kernel matrix is positive-definite, we have added the absolute of the smallest eigenvalue of the kernel matrix to the diagonal of this kernel matrix, if this smallest eigenvalue is negative (Zhang et al., 2006).

There are a number of formulations for the MKL problem, our approach builds on the simpleMKL (Rakotomamonjy et al., 2008). This formulation enables the kernel combination weights to be learnt within the SVM framework. For kernel algorithms, the solution of the learning problem of each component is of the form
\[
\hat{f}_m(x_{\text{test}}) = \sum_{i=1}^N \alpha_i K_m(x_{\text{test}}, x_{mi}) + b,
\]
where \( \alpha_i \) and \( b \) are some coefficients to be learned from samples, \( x_{\text{test}} \) is the \( m \)-th feature vector of \( x_{\text{test}} \).

In Rakotomamonjy et al. (2008), the classifier model parameter and the weights are given by the optimization problem of the MKL based on SVM as
\[
\min_{\beta_m} \sum_{m=1}^M \frac{1}{2} \| \hat{f}_m \|^2_{\| \beta_m \|} + \hat{C} \sum_{1}^{C} \xi_i,
\]
and it is restricted by \( (1) \gamma \sum_{m=1}^M f_m(x_{mi}) + y_i b \geq 1 - \xi_m \), \( (2) \xi_m \geq 0 \), and \( (4) \sum_{m=1}^M \beta_m = 1 \), where \( \xi_i \) is the slack afforded to each sample, \( \hat{C} \) is the regularization parameter, and \( \beta_m \) belongs to a different reproducing kernel Hilbert space \( \mathcal{H}_m \) associated with a kernel function. In our implementation, the stop criterion of the optimization to the formulation (8) is based on the variation of coefficient \( \beta_m \) between two consecutive steps.

**Algorithm 1.** Weight learning on facial expression recognition

input: \( M \) feature vectors \( \{ \Omega_m \}_{m=1}^M \) with \( N \) samples, \( C \) classes
output: \( \{ \beta_{m,c} \}_{m=1}^M \)

for \( c = 1 \) to \( C \) do
for \( m = 1 \) to \( M \) do
Generate Positive Set \( \Omega^+ \) — Find the data from \( \Omega_m \in c \).
Generate Negative Set \( \Omega^- \) — Find the data from \( \Omega_m \notin c \).
Form \( K_{m} = \Omega_{m} \times \Omega_{m} \).
Unit trace normalization of \( \{ K_{m} \}_{m=1}^M \).
end for
end for

Considering the discriminative information of classes, we apply the one-vs.-rest method to the formulation (8) to get the weights of the \( m \)-th feature vector under the \( c \)-th class, i.e. \( \beta_{m, c} \). In our method, for the \( c \)-th class, we firstly find the samples from \( \Omega_m \), and then we choose ones that belong to the \( c \)-th class as a positive set, and the rest are denoted as a negative set. Finally, these samples are formed as the kernel matrix \( K_{m} \). This procedure is done for all \( \{ \Omega_m \}_{m=1}^M \) under the \( c \)-th class. Based on these kernel matrices, we use the MKL (Rakotomamonjy et al., 2008) to explore the weight of the \( m \)-th feature vector to the \( c \)-th class. The procedure is summarized in Algorithm 1. In our methods, six feature vectors are generated from three facial components. The contribution of feature vectors to expressions is developed by using the MKL for weight learning. In specific case, once three feature vectors are generated by LBP or EdgeMap, the MKL can be used to observe the correlation of the components and expressions. Following the experimental setup in Section 3, we get the correlation of the components and expressions in Fig. 4. This result shows that (1) eyes play important roles in anger, contempt, and surprise; (2) mouth determines fear and sadness; (3) both nose and mouth contribute to disgust and happiness.

**5. Experiments and discussion**

The proposed approach was evaluated on the CK+ (Lucey et al., 2010) and its simulated occlusion database. We propose several methods for facial representation, occlusion detection, fusion module, and weight learning. Here, we summarize the abbreviation of our proposed methods and the comparative methods:

(a) **STLBP:** Spatio-temporal local binary pattern (Zhao and Pietikainen, 2007).
(b) **EdgeMap:** Dynamic extension of edge mapping (Gizatdinova and Surakka, 2006).
(c) **FSE:** Mean-voting fusion of STLBP and EdgeMap.
(d) **CFD:** Combination of component-based facial expression representation and fusion module.
(e) **CFD-WL:** CFD based on weight learning via the MKL.
(f) **CFD-OD:** CFD based on occlusion detection.
components, the feature-integration approach, and the feature-fusion approach, and then used LBP for three components. The recognition rates of three

Leave-One-Subject-Out method. We further divided eyes, nose, and mouth into 9 blocks, respectively, and then used LBP for three components. The recognition rates of three components, the feature-integration approach, and the feature-fusion approach are listed in Table 1. The results of each component show that mouth, eyes and nose are arranged in order of contribution to FER. Combining mouth with nose, the performance of the mouth is decreased by the feature-integration approach while kept by the feature-fusion approach. With eyes adding, the performance was reduced by the feature-integration approach, while boosted by the later one. That is, three components are affected mutually if concatenating all features into a single one. These results demonstrate that the feature-fusion approach can reduce the mutual effect of each region. Furthermore, it is observed that combining different components by the feature-fusion approach can perform better than the feature-integration approach. It means that the feature-fusion approach is an efficient and good way of handling multiple feature vectors.

5.3. Non-occluded experiments

In our approach, three components are cropped. In order to extract micro-information from components, each component is further divided into a few of the blocks. Thus, the block sizes 9 × 8, 11 × 10 and 8 × 8 are used for eyes, nose and mouth, due to their optimal performance of LBP (Huang et al., 2011). STLBP, EdgeMap, FSE, and Lucey et al. (2010) achieve the recognition rate of 87.08%, 82.77%, 92%, and 88.38%, respectively. CFD and CFD-WL achieve 89.85% and 93.23% at the recognition rate, respectively. It is observed that CFD and CFD-WL outperform the others. With occlusion detection, the performance of CFD and CFD-WL achieve 89.23% and 92.32% at the recognition rate, respectively. It is found that occlusion detection decreases the performance of CFD and CFD-WL with less than 1%. In occlusion detection, as shown in Fig. 3, the threshold is predefined. It is ideal that the reconstruction error of sub-blocks in the un-occluded facial sequence should be zero. But the reconstruction procedure may cause that reconstruction errors of some sub-blocks are higher than the pre-defined threshold. It results in that some sub-blocks might be wrong detected as occluded and the performance of CFD and CFD-WL is decreased. Even though wrongly detection happens, occlusion detection least affects FER on normal image sequences.

5.4. Experiments in the presence of eyes/mouth/lower-face facial region occlusion

In this section, we further evaluate our method, some dynamic methods (Zhao and Pietikäinen, 2007; Gizatdinova and Surakka, 2006; Kotsia et al., 2008) on more challenging environment (facial occlusion). We provide the results using CFD, CFD-OD, and CFD-OD-WL in comparison to STLBP, EdgeMap and FSE in Table 2(a)–
STLBP, EdgeMap, and FSE achieve recognition rates of (a) 63.38%, 60.62%, and 69.23% under eyes occlusion; (b) 31.38%, 24.31%, and 36.62% under mouth occlusion; (c) 10.46%, 15.69%, and 10.46% under lower-face occlusion. The recognition rates under eyes, mouth, and lower-face occlusions using CFD are 78.77%, 66.15%, and 34.46%, respectively. The experimental results show

![Fig. 5](image)

(a) Geometric structure of the simulated occlusion. (a.1) The facial image rotated with respect to the eyes positions and 68 facial points; (a.2) the region of eyes occlusion; (a.3) the region of mouth occlusion; (a.4) the region of the medical mask in lower-face region; (a.5) an example of the random patch at 50% of facial area. (b) Examples (surprise) with occlusion on eyes, mouth, lower-face regions and random patches at 5% and 50% of area of a cropped facial image from up to bottom. From left to right, it represents expression variation in the sequence. (c) Examples with different expressions with occlusions. From left to right, it represents expressions (anger, contempt, disgust, fear, happiness, sadness, surprise).

<table>
<thead>
<tr>
<th>Eyes</th>
<th>Nose</th>
<th>Mouth</th>
<th>Feature-integration approach</th>
<th>Feature-fusion approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>E + N</td>
<td>E + M</td>
</tr>
<tr>
<td>72.31</td>
<td>71.38</td>
<td>87.38</td>
<td>80.31</td>
<td>89.85</td>
</tr>
</tbody>
</table>

(c) STLBP, EdgeMap and FSE achieve the recognition rate of (a) 63.38%, 60.62% and 69.23% under eyes occlusion; (b) 31.38%, 24.31%, and 36.62% under mouth occlusion; (c) 10.46%, 15.69%, and 10.46% under lower-face occlusion. The recognition rates under eyes, mouth, and lower-face occlusions using CFD are 78.77%, 66.15%, and 34.46%, respectively. The experimental results show
that CFD is more robust to partial occlusion and has better performance than STLBP and EdgeMap. If using occlusion detection in advance, CFD achieves 85.54%, 73.54%, and 67.38% at the recognition rate under eyes, mouth, and lower-face occlusions, respectively. It is shown that occlusion detection can reduce the effect of occlusion and give an obvious improvement to CFD.

Performing FER under eyes, mouth, and lower-face occlusion by CFD-OD-WL, achieve recognition rate of 93%, 79.08%, and 73.54%, respectively.

The recognition rate on Cohn–Kanade database under eye occlusion using Gabor filters, the DNNM, and shape-based SVMs in Kotsia et al. (2008) is 86.8%, 84.2%, and 82.9%, respectively. And for lower-face occlusion, this is 84.4%, 82.9%, and 86.7%. It should be noted that the results are not directly comparable due to different experimental setups, processing methods, the number of sequences used, etc., but they still give an indication of the discriminative power of each approach. It is found that CFD-OD-WL under eye occlusion outperforms their algorithms. Unfortunately, the performance under lower-face occlusion is worse than in these methods. It is lower at about 13.16% than the best performance in Kotsia et al. (2008). The difference between their and our approach is that the of input query (image vs. video) and a different database. Different cross-validation methods are applied. Those might cause CFD-OD-WL works worse than their under lower-face occlusion.

5.5. Experiments on random occlusion patch

To prove the robustness of our approach, we extend it to the experiments on random occlusion. The experimental results based on our approach and baseline approaches are shown in Table 2(d). In Fanelli et al. (2010), the random occlusion in the video for FER is analyzed. From Fig. 6 in Fanelli et al. (2010), their experimental results are approximately from 85% to 38% with the increment of partial occlusion from 5% to 50%. From Table 2(d), with the increment of partial occlusion from 5% to 50%, the performance of STLBP is decreased sharply from 87.69% to 31.38%, as well as EdgeMap from 80.31% to 33.23% and FSE from 90.15% to 38.46%, the performance with CFD is decreased from 85.85% to 60.15%, and that of with CFD-OD is decreased from 86.15% to 76.62%, the performance with CFD-OD-WL is decreased from 90.77% to 79.69%.

Compared with STLBP, EdgeMap, FSE and Fanelli et al. (2010), the performance of CFD, CFD-OD and CFD-OD-WL is not decreased significantly. It is shown that CFD for feature representation is robust to facial occlusion even though approximately 50% of facial image is occluded. Additionally, with occlusion detection, the most of the effect of occlusion is reduced so that the CFD is improved. Although the CFD by occlusion detection is not improved significantly at 5% occlusion percentage, the performance is much improved at about 16.47% at 50% occlusion percentage. Further-

Table 2
The average recognition rate and the recognition rates of (%) all expressions using different approaches under (a) eyes occlusion, (b) mouth occlusion, (c) lower-face occlusion and (d) random occlusion, where the recognition rate of each expression is equal to the corrected expression divided by the number of samples in the expression, the average recognition rate (Av) is equal to the corrected sample in all expressions divided by the number of samples in all expressions. The bold values show the best accuracy for each expression with different occlusions.

<table>
<thead>
<tr>
<th>Method</th>
<th>An</th>
<th>Co</th>
<th>Di</th>
<th>Fe</th>
<th>Ha</th>
<th>Sa</th>
<th>Su</th>
<th>Av</th>
</tr>
</thead>
<tbody>
<tr>
<td>STLBP</td>
<td>93.33</td>
<td>11.11</td>
<td>84.75</td>
<td>32</td>
<td>68.12</td>
<td>3.57</td>
<td>69.14</td>
<td>63.38</td>
</tr>
<tr>
<td>EdgeMap</td>
<td>33.33</td>
<td>38.89</td>
<td>94.92</td>
<td>24</td>
<td>68.12</td>
<td>0</td>
<td>81.48</td>
<td>60.62</td>
</tr>
<tr>
<td>FSE</td>
<td>75.56</td>
<td>5.56</td>
<td>98.31</td>
<td>28</td>
<td>75.36</td>
<td>0</td>
<td>90.12</td>
<td>69.23</td>
</tr>
<tr>
<td>CFD</td>
<td>75.56</td>
<td>0</td>
<td>94.92</td>
<td>36</td>
<td>97.10</td>
<td>35.71</td>
<td>98.77</td>
<td>78.77</td>
</tr>
<tr>
<td>CFD-OD</td>
<td>82.22</td>
<td>38.89</td>
<td>94.92</td>
<td>48</td>
<td>100</td>
<td>60.71</td>
<td>98.77</td>
<td>85.54</td>
</tr>
<tr>
<td>CFD-OD-WL</td>
<td>91.11</td>
<td>55.56</td>
<td>96.61</td>
<td>88</td>
<td>100</td>
<td>78.57</td>
<td>98.77</td>
<td>93</td>
</tr>
</tbody>
</table>

(b) Mouth occlusion

<table>
<thead>
<tr>
<th>Method</th>
<th>An</th>
<th>Co</th>
<th>Di</th>
<th>Fe</th>
<th>Ha</th>
<th>Sa</th>
<th>Su</th>
<th>Av</th>
</tr>
</thead>
<tbody>
<tr>
<td>STLBP</td>
<td>0</td>
<td>72.22</td>
<td>98.31</td>
<td>56</td>
<td>0</td>
<td>57.14</td>
<td>1.23</td>
<td>31.38</td>
</tr>
<tr>
<td>EdgeMap</td>
<td>73.33</td>
<td>66.67</td>
<td>8.47</td>
<td>4</td>
<td>15.94</td>
<td>0</td>
<td>20.99</td>
<td>24.31</td>
</tr>
<tr>
<td>FSE</td>
<td>55.56</td>
<td>83.33</td>
<td>71.19</td>
<td>12</td>
<td>2.9</td>
<td>14.29</td>
<td>34.57</td>
<td>36.62</td>
</tr>
<tr>
<td>CFD</td>
<td>100</td>
<td>0</td>
<td>74.58</td>
<td>0</td>
<td>81.16</td>
<td>0</td>
<td>86.42</td>
<td>66.15</td>
</tr>
<tr>
<td>CFD-OD</td>
<td>75.56</td>
<td>0</td>
<td>88.14</td>
<td>0</td>
<td>100</td>
<td>17.86</td>
<td>97.53</td>
<td>73.54</td>
</tr>
<tr>
<td>CFD-OD-WL</td>
<td>82.22</td>
<td>16.67</td>
<td>86.44</td>
<td>16</td>
<td>100</td>
<td>50</td>
<td>97.53</td>
<td>79.08</td>
</tr>
</tbody>
</table>

(c) Lower-face occlusion

<table>
<thead>
<tr>
<th>Method</th>
<th>An</th>
<th>Co</th>
<th>Di</th>
<th>Fe</th>
<th>Ha</th>
<th>Sa</th>
<th>Su</th>
<th>Av</th>
</tr>
</thead>
<tbody>
<tr>
<td>STLBP</td>
<td>0</td>
<td>10.17</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>10.46</td>
<td></td>
</tr>
<tr>
<td>EdgeMap</td>
<td>0</td>
<td>50</td>
<td>1.69</td>
<td>8</td>
<td>56.52</td>
<td>0</td>
<td>0</td>
<td>15.69</td>
</tr>
<tr>
<td>FSE</td>
<td>0</td>
<td>77.78</td>
<td>8.47</td>
<td>8</td>
<td>8.7</td>
<td>25</td>
<td>0</td>
<td>10.46</td>
</tr>
<tr>
<td>CFD</td>
<td>97.78</td>
<td>0</td>
<td>33.9</td>
<td>0</td>
<td>10.14</td>
<td>0</td>
<td>50.62</td>
<td>34.46</td>
</tr>
<tr>
<td>CFD-OD</td>
<td>60</td>
<td>0</td>
<td>74.58</td>
<td>0</td>
<td>95.65</td>
<td>7.14</td>
<td>98.77</td>
<td>67.38</td>
</tr>
<tr>
<td>CFD-OD-WL</td>
<td>68.89</td>
<td>16.67</td>
<td>81.36</td>
<td>12</td>
<td>92.75</td>
<td>46.43</td>
<td>95.06</td>
<td>73.54</td>
</tr>
</tbody>
</table>

(d) Random occlusion

<table>
<thead>
<tr>
<th>Percentage of occlusion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>An</th>
<th>Co</th>
<th>Di</th>
<th>Fe</th>
<th>Ha</th>
<th>Sa</th>
<th>Su</th>
<th>Av</th>
</tr>
</thead>
<tbody>
<tr>
<td>STLBP</td>
<td>87.69</td>
<td>80.31</td>
<td>71.38</td>
<td>60.92</td>
<td>58.46</td>
<td>51.08</td>
<td>37.54</td>
<td>36.92</td>
</tr>
<tr>
<td>EdgeMap</td>
<td>80.31</td>
<td>69.85</td>
<td>62.46</td>
<td>53.84</td>
<td>48</td>
<td>47.38</td>
<td>46.15</td>
<td>38.46</td>
</tr>
<tr>
<td>FSE</td>
<td>90.15</td>
<td>88</td>
<td>78.77</td>
<td>72.23</td>
<td>65.85</td>
<td>63.08</td>
<td>53.54</td>
<td>49.85</td>
</tr>
<tr>
<td>CFD</td>
<td>85.85</td>
<td>82.15</td>
<td>80.31</td>
<td>79.69</td>
<td>76.92</td>
<td>74.77</td>
<td>68</td>
<td>65.23</td>
</tr>
<tr>
<td>CFD-OD</td>
<td>86.15</td>
<td>84.62</td>
<td>81.23</td>
<td>82.46</td>
<td>82.15</td>
<td>80</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td>CFD-OD-WL</td>
<td>90.77</td>
<td>90.15</td>
<td>88</td>
<td>87.68</td>
<td>86.77</td>
<td>84</td>
<td>83.69</td>
<td>81.85</td>
</tr>
</tbody>
</table>
more, CFD-OD obtains an improvement by weight learning. These results therefore show that the proposed framework CFD-OD-WL has significant robustness to the common and random occlusions.

5.6. Experimental analysis

Overall, for all facial expressions, non-occluded and occluded experiments are conducted as a simulation of FER in an uncontrolled environment. Fig. 5(c) shows the seven expressions under three occlusions. In Fig. 5(c), we can easily recognize most of the expressions under eyes occlusion through the mouth. Table 2(a) shows most of methods can recognize most of the expression under eyes occlusion. It is moderate to recognize surprise and happiness by human or machine. The following ones are then anger and fear. But contempt and sadness are not easy to recognize through the mouth by the machine. With lower-face occlusion, the performances of most of the expressions are sharply decreased compared with eyes and mouth occlusions. It means that the expressions under lower-face occlusion are difficult to be recognized by eyes, because the majority of discriminative information is in lower-face region. Unfortunately, the lower-face occlusion always exists in our lives, for example, we wear medical mask in the hospital or scarf in the winter. That is, lower-face occlusion poses a serious challenge to FER. In Section 5.4, we analyze other methods (STLBP, EdgeMap and FSE) under three occlusions. The comparative study shows that CFD-OD-WL can perform well on occlusion for FER. Furthermore, CFD-OD-WL obtains the best performance on (1) contempt, fear, happiness, sadness and surprise under eyes occlusion, (2) happiness and surprise under mouth occlusion, and (3) disgust and fear under lower-face occlusion. Moreover, even though the occlusion detection decreases the performance of CFD and CFD-WL in the non-occluded experiment, we further prove that the occlusion detection provides an improvement in recognizing expressions under partial occlusion. Comparing with CFD and CFD-OD under eyes occlusion, the occlusion detection improves the performance of CFD on most of the expressions. Thereafter, it appears to be a similar trend towards the mouth and lower-face occlusions. The comparative results show that the occlusion detection serves a positive and effective function to FER. Considering that the random occlusion appears to be more available in the video, through the complementary experiment, the performance of our system was increasing less than for other methods. Comparing to the state-of-the-art method (Fanelli et al., 2010), it is found that occlusion detection is appropriate to FER.

6. Conclusion

In this paper, we proposed a framework of integrating the component-based feature descriptors, a sparse-representation based occlusion detection, and a weight learning to recognize facial expression from video sequences, especially under partial occlusion. The proposed method can automatically detect in the target video the presence of occlusion and remove it if needed. In order to be robust to partial occlusion, dynamic appearance representation and statistical shape representations for three components (eyes, nose and mouth) are extracted by STLBP and EdgeMap, respectively. The component-based approach can provide more discriminative information and be robust to partial occlusion. Concerning the effect of facial occlusion, the sparse representation is used in the occlusion detection for detecting the occluded part in three components. Occlusion detection can mostly reduce the effect of occlusion. Finally, we proposed the use of the MKL to learn the weight of components for boosting the performance after occlusion detection. Challenging sets of simulated occlusion experiments demonstrate that the proposed approach achieves high-recognition accuracy on different occlusions. It shows that the component-based approach and weight learning can boost the performance of FER. Occlusion detection is also proved to be reasonable and effective.

The component-based feature extraction, including multiple feature vectors, makes our approach robust to partial occlusion. The usefulness of occlusion detection was justified for FER under different occlusions in Section 5. It would be interesting to see how the proposed system would perform in comparison to state-of-the-art descriptors in handling partial occlusion. In the case where occluded parts are removed, we may be able to boost the facial expression recognition using weight learning. Toward partial occlusion, we could benefit from the proposed approach. Future research includes the occlusion of random frames in the videos, as well as the application of the algorithm in other facial expression databases.

Acknowledgement

This work was supported by the Academy of Finland and Infotech Oulu. The support of the Finnish Agency for Technology and Innovation (Tekes) is also gratefully acknowledged. This work was partly supported by the National Basic Research Program of China (973 Program) under Grant No. 2011CB302202, and partly by NSFC under Grant No. 61073137 and Jiangsu NSF under Grant No. BK2010243.

References