Multimed Tools Appl (2018) 77:917–937
DOI 10.1007/s11042-016-4321-2

Evaluating real-life performance of the state-of-the-art in facial expression recognition using a novel YouTube-based datasets

Muhammad Hameed Siddiqi[1] · Maqbool Ali[2] · Mohamed Elsayed Abdelrahman Eldib[3] · Asfandyar Khan[4] · Oresti Banos[5] · Adil Mehmood Khan[6] · Sungyoung Lee[2] · Hyunseung Choo[1]

Received: 17 January 2016 / Revised: 31 October 2016 / Accepted: 27 December 2016 /
Published online: 6 January 2017

Abstract Facial expression recognition (FER) is one of the most active areas of research in computer science, due to its importance in a large number of application domains. Over the years, a great number of FER systems have been implemented, each surpassing the other in terms of classification accuracy. However, one major weakness found in the previous studies is that they have all used standard datasets for their evaluations and comparisons. Though this serves well given the needs of a fair comparison with existing systems, it is argued that this does not go in hand with the fact that these systems are built with a hope of eventually being used in the real-world. It is because these datasets assume a predefined camera setup, consist of mostly posed expressions collected in a controlled setting, using fixed background and static ambient settings, and having low variations in the face size and camera angles, which is not the case in a dynamic real-world. The contributions of this work are two-fold: firstly, using numerous online resources and also our own setup, we have collected a rich FER dataset keeping in mind the above mentioned problems. Secondly,

✉ Hyunseung Choo
choo@skku.edu

Muhammad Hameed Siddiqi
siddiqi@skku.edu

[1] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea

[2] Department of Computer Engineering, Kyung Hee University, Suwon, Korea

[3] Department of Biomedical Engineering, Kyung Hee University, Suwon, Korea

[4] Department of Computer Science, University of Science & Technology, Bannu, Pakistan

[5] Center for Telematics and Information Technology, University of Twente, Enschede, 7500AE, Netherlands

[6] Department of Computer Science, Innopolis University, Kazan, Russia

we have chosen eleven state-of-the-art FER systems, implemented them and performed a rigorous evaluation of these systems using our dataset. The results confirm our hypothesis that even the most accurate existing FER systems are not ready to face the challenges of a dynamic real-world. We hope that our dataset would become a benchmark to assess the real-life performance of future FER systems.

Keywords Facial expressions · Classification · YouTube · Real-life scenarios

1 Introduction

Knowledge about people's emotions can serve as an important context for automatic service delivery in a large number of context-aware systems. Many research applications of image processing and pattern recognition, such as human computer interaction [3], robot control and driver state surveillance [49], and human behavior studies in telemedicine and e-health environments [24], can benefit from the knowledge of people's emotions. Hence, human facial expression recognition (FER) has emerged as an important research area over the last two decades.

Facial expression recognition can be classified into two categories: First is pose-based FER [28, 47], which deals with recognizing artificial expressions: expressions produced by people when they are asked to do so [5]. The second is spontaneous FER [4, 50], which deals with the expressions that people give out spontaneously, and these are the ones that can be observed on a day-to-day basis, such as during conversations or while watching movies [5].

A typical FER system consists of four main sub-components: preprocessing, feature extraction, feature selection, and recognition modules. In preprocessing, the image quality is improved, and the faces are located in the expressions frames before recognizing the expressions. Feature extraction deals with extracting the distinguishable features from each facial expression shape and quantizing them as discrete symbols. Feature selection is used for selecting a subset of relevant features from a large number of features extracted from the input data. Finally, in recognition, a classifier is first trained using the training data, which is then used to generate labels for the expressions in the incoming video data [37].

A great deal of research effort has gone into designing efficient and accurate FER systems in the past, and a variety of techniques for each component has been proposed [1, 16–18, 31, 33], which will be discussed later. One major weakness with almost all of the state-of-the-art approaches in FER, including our own system [36], is the way those systems have been evaluated. Every FER system is designed with a motivation to be used in a real-life scenario; however, when it comes to testing and validating the recognition performance of these systems, standard datasets are employed for both training and testing. Though it serves well for the sake of comparison with existing approaches, which were also tested using the same datasets, such results cannot be used as a representative of an FER system's performance in real-life. It is because almost all of these datasets were collected using specific kinds of video cameras, which might not be the case in real world. Furthermore, a majority of these datasets was collected in controlled environments under constant ambient settings and did not take into account the color features and factors such as gender, race, and age. Some of the previous datasets did not consider whether the subjects wore glasses or if they had a beard. Another important element in FER domain is the size of a subject's face that, in real life, can vary from person to person. Also, it can differ depending on how far the subject is from the camera. However, in most of the previously used datasets, the face size did not vary much, mainly due to a predefined setup of the cameras. All previous datasets were collected either indoors or outdoor under static scenarios. In most of the datasets, the expressions were recorded mainly from the frontal view of the subjects, with only a slight variation, which might not be the case in a dynamic real world.

In short, existing FER systems utilized publicly available datasets and did not consider the real world challenges in their respective systems. Since the beginning of research in FER, the focus has been on designing new and improved methodologies, and evaluating them using publicly controlled-settings datasets for the sake of a fair comparison. Little or no effort has been put into designing a new dataset that is closer to real-life situations, probably because creating such a dataset is a very difficult and time consuming task. Accordingly, this work makes the following contributions:

- We have defined a comprehensive, realistic and innovative dataset collected in-house as well as from online sources, such as YouTube, for real-life evaluation of FER systems. From indoor lab settings to real-life situations, we collected three cases with increasing complexity. In the first case, ordinary subjects performed expressions in a pose-based manner, with dynamic background, lighting and camera settings. Hence, these expressions are pose-based expressions in an uncontrolled environment. In the second case, the expressions were collected from the movie/drama scenes of professional actors and actresses. Though these are also pose-based expression, we had no control on expression production, camera, lighting and background settings. Hence, these expressions are semi-naturalistic expressions under dynamic settings. Finally, in the third case, the expressions were recorded from real world talk shows, news, and interviews. Hence, these expressions are spontaneous expressions collected in natural and dynamic settings. In all three cases, a large number of different subjects of different gender, race, and age were included. Also, many subjects wore glasses and had a beard.
- From the existing work in FER, more recent and highly accurate FER systems including [1, 16–18, 20, 29–31, 33–35] were selected and implemented.
- After implementation, all these systems were tested on the three collected datasets, and a detailed analysis of their performance was produced and presented in this paper.
- Based on the obtained results, components are identified that are crucial to a satisfactory performance of an FER system in real-life situations.

The rest of the paper is organized as follows. Section 2 reviews the existing standard datasets of facial expression and recent published FER systems. Section 3 describes the defined datasets. The experimental setup, results, and discussion are presented in Section 4. Finally, the paper concludes with future directions in Section 5.

2 Related works

2.1 Existing datasets

Table 1 provides a short but thorough review of previously used datasets for evaluating the performance of existing FER studies. We can see that the most of the datasets were collected either indoors or outdoors, in controlled conditions under identical ambient settings with fixed or similar backgrounds. These assumptions can not be held true in the dynamic real world. Furthermore, when recording expression, variations in gender, age, race and color were not taken into account. Even in the studies where multiple subjects were considered, having different age, race, and gender; the face size did not vary much as subjects were at the same distance from the camera. Furthermore, other facial features like wearing glasses,

Table 1 Summary of the existing publicly available standard datasets of facial expressions and their limitations

Dataset Name	Source	Characteristics	Limitations
FERET [42]	2 RGB Cameras	2,413 still images including 1,199 individuals and 365 duplicate sets of images	Pose-based dataset with frontal images collected under a semi-controlled environment with predefined setup of camera and background to maintain a degree of consistency
SCface [14]	5 Video Surveillance Cameras	4,160 still images from 130 subjects (male and female) in indoor environment	Pose-based dataset with frontal images collected under a predefined camera setup and a constant background
CMU-PIE [38]	RGB Camera	41,368 images of 68 people (male and female) with 13 different poses, 43 different illumination conditions, and 4 different expressions	Pose-based dataset collected in indoor environment having only a few expressions with frontal view of the camera
Multi-PIE [15]	RGB Camera	750,000 images collected from 337 subjects under 15 view points and 19 illumination conditions	Pose-based dataset with only frontal camera view, collected under specific ambient conditions
Yale Face B [13]	Single RGB Camera	5,760 images of 10 subjects under 576 viewing conditions (9 poses x 64 illumination conditions) with a single light source and static ambient (background) illumination	Pose-based dataset with images from the frontal view; under a predefined setup of camera, light, and background
AT&T [32]	Video Camera	40 distinct subjects (10 different images each) taken against a dark homogeneous background with the subjects in an upright, frontal position with varying the lighting, having minimal expressions such as open/closed eyes, smiling/not smiling, glasses/no glasses	Pose-based dataset that was collected in indoor environment under predefined setup of camera and background. Also, the images are from the frontal view
Cohn-Kanade (CK) [21]	Video Camera	486 sequences from 97 persons (university students) having the age range from 18–30 and most of them were female. The dataset has seven basic expressions	Pose-based dataset collected under constant light and background with frontal view of the camera
Extended CK+ [25]	Video Camera	Solved the problems of CK dataset; has 593 video sequences including pose-based and spontaneous expressions from 123 subjects	The dataset was collected under a controlled environment with constant background. Though some subjects were at a 30-degree angle with the camera, the remaining subjects were with frontal view of the camera.

Table 1 (continued)

Dataset Name	Source	Characteristics	Limitations
NIST MID [12]	Kodak MegaPixel camera	3,248 images collected from 1,495 male and 78 female subjects	Limited number of images from the side views; collected in a controlled environment with a static background.
AR Face [2]	RGB Camera	4,000 color images from 126 subjects (70 men and 56 women) with frontal view faces and different facial expressions, illumination conditions, and occlusions (sun glasses and scarf)	Pose-based dataset that was collected under a controlled environment with predefined setup of camera and static background
Oulu Physics-Based Face [27]	RGB Camera	125 people performed the expression (16 pictures per person) with 16 different camera calibration and illuminations	Pose-based dataset that was collected under a dark room conditions with specific RGB cameras under a controlled environment with a constant background
JAFFE [26]	Video Camera	213 images of 7 facial expressions (6 basic facial expressions and 1 neutral) posed by 10 Japanese female	All of the images were taken from the frontal view of the camera with tied hair in order to expose all the sensitive regions of the face, under strictly controlled environment with static background and the subjects were only female
BioID Face [6]	Video Camera	1,521 images from 23 persons	Pose-based dataset with only a few expressions from only the frontal view of the camera with constant light and predefined setup of the camera and background
Georgia Tech Face [7]	RGB Camera	50 people performed the expressions in two or three sessions and the images are represented by 15 color JPEG images with cluttered background	Pose-based dataset with predefined setup of the camera
Indian Face [19]	RGB Camera	40 distinct subjects (11 images of each) performed different expressions that were taken against a bright homogeneous background with the subjects in an upright, frontal position	Pose-based dataset with a small number of expressions, captured using a static background under a controlled environment with fixed camera settings
Labeled Faces in the Wild [46]	World Wide Web	More than 13,000 images collected from the web. Among them, 1,680 of the people pictured have two or more distinct photos in the database	Spontaneous dataset that was only collected from the web and a commercial face alignment software was utilized in order to align the faces

Table 1 (continued)

Dataset Name	Source	Characteristics	Limitations
UCD Colour Face Image [41]	Video Camera, Scanner, World Wide Web	100 color images of faces (in indoors, outdoors, complex, simple) in which the subjects have beards, mustaches, and glasses	Spontaneous dataset that has a limited number of expressions and most of them were from frontal view of the camera with low quality resolution due to ambient noise
FRAV2D [9]	Video Camera	3,488 image from 109 subjects (75 men and 34 women) with 32 images per person	Pose-based dataset with predefined setup of the camera. Most of the images were from the frontal view, collected against a plain and dark blue background
PUT Face [23]	Video Camera	9,971 images of 100 people and the images of each were taken in five series with different orientation	Pose-based dataset with a limited number of expressions that were collected in controlled ambient conditions with a uniform background and a fixed camera settings
Plastic Surgery Face [39]	Video Camera	1,800 pre and post surgery images from 900 subjects (For each subject, there are two frontal face images with proper illumination and neutral expression)	Spontaneous dataset that has a limited number of expressions with frontal view of the camera. The images were taken under predefined ambient conditions, camera settings, and background
PolyU NIR Face [48]	JAI Camera	35,000 image collected from 350 subjects (100 images per person) and the subjects were asked to move near to or away from the camera in a certain range	Pose-based dataset with images from the frontal view of the camera. The dataset was collected in a controlled environment with predefined setup of camera, light, and background
USTC-NVIE [44]	Infrared Camera	108 subjects (university students) performed both pose and spontaneous expressions, and the age range for the subjects was from 17 to 31 years	An infrared thermal camera was used for dataset collection with a predefined lighting setup and background
FEI Face [43]	Video Camera	2,800 images were collected from 200 individuals (14 images for each person) with different angle orientations	Pose-based dataset that was collected against a white homogeneous background with fixed camera settings. The images are in an upright frontal position with profile rotation of up to about 180 degrees
VADANA [40]	Digital Camera	2,298 images from 43 subjects (53 images per person) with wide range of variations in pose, expression, illumination, distractions such as spectacles, facial hair, and partial occlusions	Pose-based dataset having most of the images from the frontal view of the camera with predefined setup of the camera and background

Table 1 (continued)

Dataset Name	Source	Characteristics	Limitations
LDHF-DB [22]	Infrared Camera	100 subjects were involved (70 males and 30 females) for data collection at distances of 60m, 100m, and 150m outdoors and at a 1m distance indoors	Pose-based dataset collected with a static background under a predefined setup of the camera and light and the subjects were trained before performing the expressions
YouTube Faces [45]	YouTube	3,425 videos of 1,595 different people (An average of 2.15 videos are available for each subject). The shortest clip duration is 48 frames, the longest clip is 6,070 frames, and the average length of a video clip is 181.3 frames	Spontaneous datasets in which the images are only collected from the performances of actors from YouTube, which is not naturalistic, and all the images are manually verified in order to check whether the subjects are correctly labeled or not
YMU [8]	YouTube	604 image from 151 subjects, 4 images per subject (2 before and 2 after the makeup)	Spontaneous dataset collected from only female subjects

having a beard and keeping different hairstyles were mostly ignored. Finally, in most of the datasets, the expressions were recorded mainly from the frontal view of the subjects, with only a slight variation, which might not be the case in real life.

2.2 Existing FER systems

Similar to Table 1, Table 2 provides a summary of the existing FER systems. Mainly, those techniques are discussed that have shown a high accuracy when evaluated using the existing datasets, and which we were able to implement. For each system, Table 2 provides the methodology (feature extraction, feature selection, and classification), the dataset used for evaluation, and the recognition accuracy achieved on that dataset.

3 Proposed dataset

As stated earlier, the main aim of this research was to collect a unique and comprehensive dataset that any FER system can employ to evaluate its real performance for identifying the desired emotions correctly and efficiently from a variety of subjects from across the globe. When collecting this dataset, limitations of the existing datasets were considered, and a significant amount of time was spent on selecting the most appropriate images with relevant emotions, situation, and surroundings. In total, three sub-datasets were collected: emulated, semi-naturalistic, and naturalistic datasets. Each dataset contains six basic expressions: happy, sad, angry, normal, disgust, and fear. The description of each of these datasets is as follows.

- *Emulated Dataset*: Emulated dataset is a mixture of front-faced images collected from the existing pose-based facial expression datasets, and the pose-based expressions collected in-house using our own testbed. For the latter case, 50 subjects (male: 25, female:

Table 2 Performance summary of the existing FER systems on publicly available standard datasets of facial expressions (Unit: %)

FER Systems	Methodology	Used Datasets	Results
AH-ASM [33]	Haar-like features with SVM	CK+	88%
W-BPNN [1]	Haar, Daubechies, and Coiflet wavelets with neural network	JAFFE	94%
LDN-SVM [31]	Local directional number pattern (LDN) with SVM	CK+, MMI, CMU	92%
CLM-SVM [16]	Discriminative response map fitting (salient patches) with SVM.	CK+, JAFFE	92%
RLBP-NN [29]	Robust local binary pattern (RLBP) with curvelet transform and nearest neighbor (NN)	JAFFE	97%
PHOG-SVM [17]	Pyramid histogram of gradients (PHOG) with LBP and SVM	JAFFE, CK+	93%
CNF-FER [34]	Curvelet transform with normalized mutual information and HMM	CK, JAFFE, Yale B, USTC-NVIE	98%
LDP-SVM [18]	Local directional pattern (LDP) with PCA and SVM	CK, JAFFE	96%
LDPv-SVM [20]	Local directional pattern variance (LDPv) with SVM	CK	94%
LBP-SVM [30]	LBP patches with multi-layer model	CK	92%
OLDA-HMM [35]	Optical flow with LDA and HMM	CK, JAFFE, USTC-NVIE, Yale B, FEI	98%

25, aged between 20 - 35 years old) were hired to perform each of the six targeted expressions. For each expression, we collected over 165 images in our lab under varying ambient settings and changing background. The images used in this dataset are of the size 240×320 and 320×240 pixels. Six sample images from this dataset are shown in Fig. 1.

- *Semi-naturalistic Dataset*: To construct this dataset, we downloaded and thoroughly watched hundreds of online available movies, videos, and shows from various sources including YouTube, Dailymotion, and other online available media sources. The selection of source videos was made such that the subjects in them are from across the globe (actors and actresses from the Hollywood, Bollywood, and Lollywood). Furthermore, they belong to a variety of ethnicities (Asian, American, African, European, etc.); age groups (4 to 60 years old); gender (male and female); and have varying facial structural properties (such as with/without beard).

  Moreover, from each video we chose images that represented real life scenarios and contributed to the benefit of the dataset for evaluation and efficiency. For example, we collected images with different facial orientations, such as frontal, right-sided, left-sided, etc. The videos were in high definition quality, and the images were separately extracted using an image capturing software called GOMPlayer software [10] that is

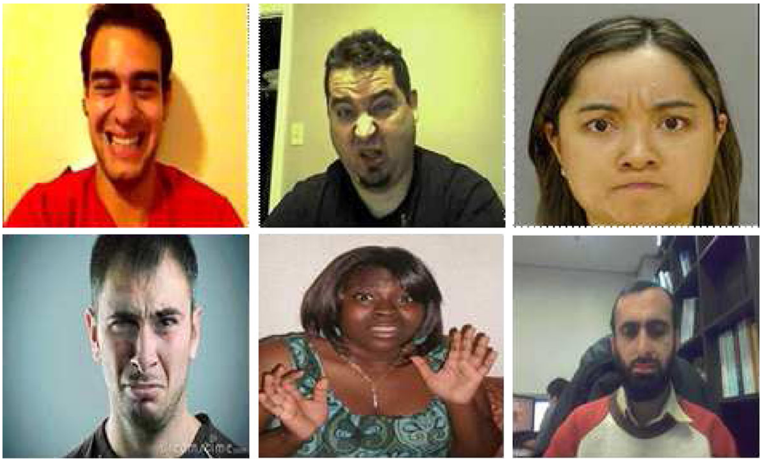


Fig. 1 Sample images (*happy*, *anger*, *sad*, *disgust*, *fear*, and *normal*) from the emulated dataset

freely available online and captures images in user defined resolution and image quality. The generated images are all in ".jpg" format, whereas the videos were in ".avi" format. Similar to emulated dataset, each expression has over 165 images in this dataset, too. The image size is 240×320 and 320×240 pixels. Six sample images from this dataset are shown in Fig. 2.

– *Naturalistic Dataset*: Unlike other datasets, we collected the naturalistic dataset purely from the talk shows, interviews, and other natural videos (such as news and recordings of real life incidents). Such a source makes this dataset more vibrant and suitable for real life testing of an FER system. To collect this dataset, we went through a tough situation of selecting appropriate emotions and capturing them at the right time, with the right mood. Just like the semi-naturalistic dataset, the subjects in this dataset do not represent a particular community class. They belong to various parts of the world, race, age (10 to

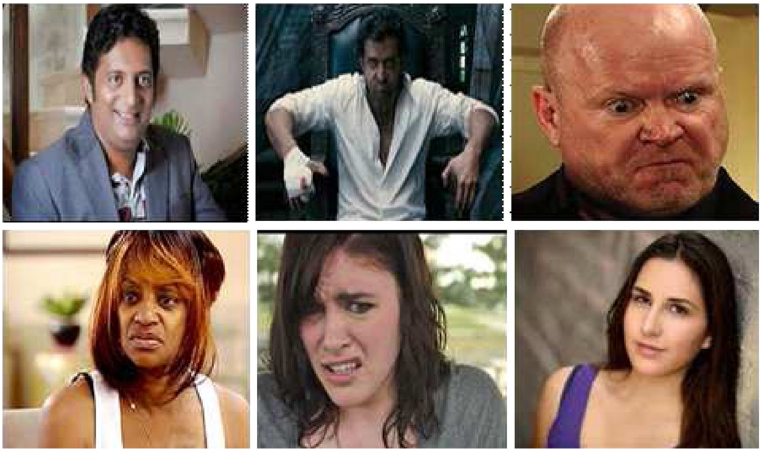


Fig. 2 Sample images (happy, anger, sad, disgust, fear, and normal) from the semi-naturalistic dataset

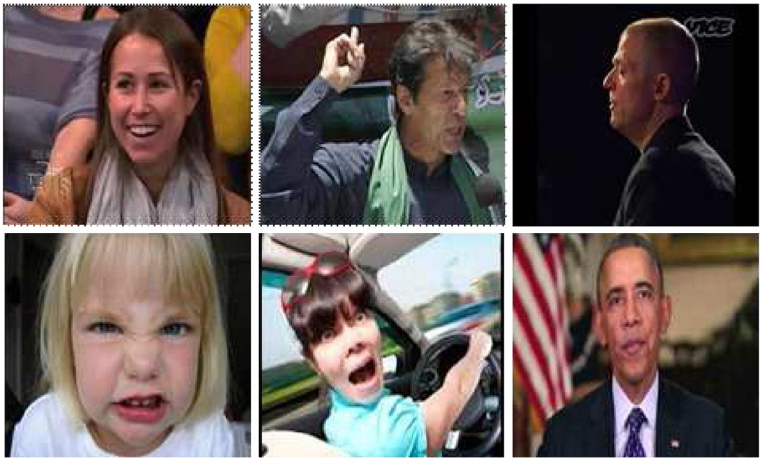


Fig. 3 Sample images (happy, anger, sad, disgust, fear, and normal) for the naturalistic dataset

50 years old), and gender. However, unlike semi-naturalistic dataset the subjects in this dataset are not actors and include doctors, patients, politicians, instructors to children, and workers, etc.

Similar to the semi-naturalistic dataset, images in this dataset reflect real life situations. These include a variety of backgrounds, unintentional expressions of the subjects, expressions from different facial orientations, and both indoors and outdoor locations under different ambient settings, etc. Moreover, subjects with/without glasses, open/closed hair, with/without a hat, and other complex scenarios were considered. For each expression, over 165 images were collected. Similar to the other two datasets, the images used in the dataset are of size 240×320 and 320×240 pixels. Six sample images from the naturalistic dataset are shown in Fig. 3.

The collection of datasets began in September 2014 and finished in February 2015. GOMPlayer software was used for capturing the images from the videos. All images were resized by using Fotosizer software [11] in order to bring a consistency among the expression images. These datasets are made publicly available at (https://github.com/hameedsiddiqui/dataPublic.git) for the research community.

4 Experimental results and discussion

4.1 Experimental setup

The eleven FER techniques, listed in Table 2, were implemented and tested on the collected datasets in a set of two experiments. Each of these experiments was performed in Matlab using an Intel Pentium Dual-Core^{TM} (2.5 GHz) with a RAM capacity of 3 GB. A brief description of the experiments is given below.

- In the first experiment, we used the $10-$fold cross-validation rule to measure the recognition accuracy of each FER system for the three datasets. In other words, each dataset was divided into ten random subsets. Out of these ten subsets, one subset was used as the validation data, whereas the remaining nine subsets were used as the training data, and this process (training and testing) was repeated ten times, each time picking a new subset as the validation data. The overall process, division into random sets and applying the 10−fold cross-validation, was repeated 20 times.

- On the other hand, in the second experiment $n-$fold cross-validation scheme was applied based on datasets. In other words, from the three datasets, two were used as validation data, whereas the remaining one dataset served as the training data. This process was repeated three times, with data from each dataset used exactly once as the training data.

4.2 Experimental results

Table 3 provides the results (recognition accuracy and standard deviation) obtained by each FER technique in both experiments. It also gives a breakdown of each FER technique concerning its architectural elements.

4.2.1 Overall analysis

It can be seen in Table 3 that in the first experiment majority of the systems showed a good performance (within the range 70 to 90 %) on emulated dataset. Their performance dropped by 10 to 15 % on semi-naturalistic scenarios, and as expected, their performance was dramatically reduced between 20 to 25 % on the naturalistic dataset.

In the second experiment when all the systems were trained using the emulated dataset and tested on naturalistic and semi-naturalistic datasets, the recognition accuracy of each system is much less than their respective accuracies, where all these systems were trained using the same emulated dataset; however, testing was done using the samples from the emulated dataset, too. This shows that an FER system that has achieved very high recognition accuracy for pose-based dataset, collected in controlled settings, cannot be expected to yield the same high accuracy when deployed to be used in the real-world.

The performance of all the systems went further down by 19 % when trained on semi-naturalistic datasets and tested on the emulated and naturalistic datasets; and by 27 % when trained on naturalistic dataset and tested on emulated and semi-naturalistic datasets. This clearly tells us that the FER systems, even the ones that have provided impressive results for the standard datasets, are not yet ready to handle the challenges of a highly dynamic real-life scenario. These challenges include: subjects with different facial features, gender, race, and age; varying lighting conditions; high variations in angle to the camera, difference in size of the face that is related to proximity. These are only some of the factors that can cause misclassification.

4.2.2 Detailed analysis

Among the eleven FER systems implemented and tested in this work, CNF-FER and OLDA-HMM showed better performance on previous datasets, as well as on the proposed datasets. CNF-FER reported the recognition accuracy of 98 % on existing dataset (as indicated in Table 2). As for the proposed datasets, we observed recognition accuracy of about 90 % on emulated, 78 % on semi-naturalistic, and 73 % on naturalistic datasets (as shown in Table 3). We believe that the reported high accuracy and an acceptable performance on the proposed datasets is because CNF-FER employs a feature selection method on top of curvelet transform in the frequency domain. The feature selection is performed using normalized mutual information criteria based on max-relevance and min-redundancy (mRMR)

Table 3 Architectural breakdown of the eleven FER techniques and their performance (recognition accuracy and standard deviation) in both experiments

FER System	Preprocessing	Feature Extraction		Feature Selection	Classification		Results: Accuracy (%) ± Standard Deviation	
		Single Features	Hybrid Features		Frame Based	Sequential Based	First Experiment	Second Experiment
AH-ASM	✗	✓	✗	✗	✓	✗	ED: 70 ±5.2 SND: 62 ±2.4 ND: 50 ±4.5	Training on ED: 59 ±3.6 Training on SND: 40 ±1.9 Training on ND: 32 ±2.9
W-BPNN	✗	✗	✓	✗	✓	✗	ED: 77 ±3.1 SND: 71 ±1.9 ND: 62 ±3.1	Training on ED: 62 ±2.9 Training on SND: 51 ±3.5 Training on ND: 40 ±3.7
LDN-SVM	✗	✓	✗	✗	✓	✗	ED: 87 ±2.7 SND: 73 ±1.8 ND: 64 ±2.9	Training on ED: 66 ±4.6 Training on SND: 51 ±2.5 Training on ND: 31 ±5.0
CLM-SVM	✗	✓	✗	✓	✓	✗	ED: 88 ±1.1 SND: 71 ±3.9 ND: 60 ±3.5	Training on ED: 64 ±0.8 Training on SND: 49 ±3.4 Training on ND: 30 ±2.7
RLBP-NN	✗	✗	✓	✗	✓	✗	ED: 73 ±2.9 SND: 67 ±3.1 ND: 64 ±1.7	Training on ED: 65 ±4.1 Training on SND: 47 ±2.7 Training on ND: 42 ±2.1
PHOG-SVM	✓	✗	✓	✓	✓	✗	ED: 75 ±4.2 SND: 70 ±2.5 ND: 59 ±5.6	Training on ED: 66 ±1.0 Training on SND: 49 ±3.6 Training on ND: 45 ±4.5
CNF-FER	✗	✓	✗	✓	✗	✓	ED: 90 ±4.8 SND: 78 ±2.9 ND: 73 ±3.0	Training on ED: 70 ±5.2 Training on SND: 50 ±2.2 Training on ND: 47 ±3.6
LDP-SVM	✗	✓	✗	✓	✓	✗	ED: 88 ±2.3 SND: 75 ±1.0 ND: 67 ±2.7	Training on ED: 68 ±4.0 Training on SND: 48 ±1.9 Training on ND: 45 ±4.4

Table 3 Architectural breakdown of the eleven FER techniques and their performance (recognition accuracy and standard deviation) in both experiments (continued)

FER System	Preprocessing	Feature Extraction		Feature Selection	Classification		Results: Accuracy (%) ± Standard Deviation	
		Single Features	Hybrid Features		Frame Based	Sequential Based	First Experiment	Second Experiment
LDPv-SVM	✗	✓	✗	✓	✓	✗	ED: 81 ±2.4 SND: 70 ±4.2 ND: 57 ±2.6	Training on ED: 56 ±3.8 Training on SND: 45 ±3.3 Training on ND: 37 ±3.8
LBP-SVM	✗	✓	✗	✗	✓	✗	ED: 78 ±3.1 SND: 72 ±3.5 ND: 61 ±3.5	Training on ED: 54 ±2.7 Training on SND: 44 ±4.7 Training on ND: 40 ±1.7
OLDA-HMM	✓	✓	✗	✓	✗	✓	ED: 91 ±2.0 SND: 80 ±2.4 ND: 72 ±4.8	Training on ED: 72 ±3.0 Training on SND: 58 ±4.1 Training on ND: 39 ±2.2

Here, ED stands for emulated dataset, SND stands for semi-naturalistic dataset, and ND stands for naturalistic dataset methods, which helps the system in getting rid of unnecessary features and improves the overall feature space. Similarly, OLDA-HMM reported 98 % accuracy on existing datasets (as indicated in Table 2), and gave 91 % on emulated, 80 % on semi-naturalistic, and 72 % on naturalistic datasets (as shown in Table 3). The facial features are very sensitive to noise and illumination changes. OLDA-HMM uses a preprocessing method to minimize such noise. Moreover, it also employs a feature selection method, based on the forward selection and backward regression model, to remove the unnecessary features. It is due to these factors that OLDA-HMM got a high original recognition accuracy, and showed an adequate performance on the proposed datasets, too. Finally, both CNF-FER and OLDA-HMM use a sequence-based classifier, which enables them to use temporal information for a better performance.

On the other hand, LDN-SVM, CLM-SVM, LDP-SVM, and LDPv-SVM showed better performance on emulated and semi-naturalistic datasets (as shown in Table 3). LDN-SVM got 87 % on emulated and 73 % on semi-naturalistic datasets. CLM-SVM got 88 % on emulated and 71 % on semi-naturalistic. LDP-SVM attained 88 % on emulated and 78 % on semi-naturalistic datasets. LDPv-SVM got 81 % on emulated and 70 % on semi-naturalistic datasets. However, the results were not as satisfactory when these methods were applied to the naturalistic dataset. LDN-SVM achieved 64 %, CLM-SVM got 60 %, LDP-SVM attained 67 %, and LDPv-SVM got only 57 % recognition accuracy. It could be because all of these FER systems extract local features. Furthermore, they do not employ the preprocessing step. As a result, the features extracted by these systems get affected by the dynamic backgrounds, changing ambient settings, and other variations that are present in the naturalistic dataset. Finally, all of these systems use frame-based classification, which relies on extracting information from only the current frame.

Next, W-BPNN and LBP-SVM showed better performance only on emulated dataset (as shown in Table 3). This is because these systems are specifically designed for the indoor environment and do not possess the ability to show better performance in outdoor settings. Thus, their performance degraded to a great extent when applied to semi-naturalistic and naturalistic datasets.

Finally, AH-ASM did not show a satisfactory performance on any of the proposed datasets (as shown in Table 3). This is because the system uses active shape model with Haar-like features. Under these settings, some specific intensity values are used that can vary in different scenarios and thus can cause misclassification.

5 Conclusion and future direction

A significant number of very accurate and efficient FER systems have been proposed over the last decade, which have yielded high recognition accuracies when tested on existing standard FER datasets. However, this does not guarantee them displaying the same performance in real-world situations. It is because the existing datasets collected facial expressions under a predefined setup and camera deployment. It is an assumption that cannot hold true in real-life scenarios. Furthermore, these datasets are mostly pose-based and were collected in a controlled environment with constant background and ambient conditions.

Accordingly, in this work, we have compiled a rich FER dataset, which consists of three sub-datasets: emulated, semi-naturalistic, and naturalistic datasets. We put our utmost effort into making sure that the datasets we collected would closely represent the real-world. They consist of a vast number of subjects of different gender, race, and age. Instead of using a fixed settings, the datasets were collected from various situations having different backgrounds, proximity to the camera (it affects the size of the face), camera angles, ambient settings, and ambient noise. Subjects have different facial features, too such as glasses and beard.

Also, we implemented eleven state-of-the-art FER systems and evaluated their performance using our datasets in a set of two experiments. Based on the experimental results we conclude the following.

- The facial features are very sensitive to noise and changes in ambient settings. These factors can frequently change in the real life. Therefore, it is essential for FER systems to have a preprocessing method to handle such noise to cope with the challenges of the dynamic real world.
- Several parts of a human face contribute towards expressions making, and extracting features from these parts can help FER systems to classify the expressions accurately. However, relying only on a single type of features won't suffice in real life situations, and thus, hybrid feature extraction techniques should be explored.
- Even after proper and efficient feature extraction, there might be some redundancy among the features. Therefore, a feature selection method is advised to select only the most informative features and remove unnecessary features from the feature space.
- Using a frame-based classification limits FER systems to using only the current frame without any reference image (neutral face image). This results in loss of information, which may cause misclassification. Therefore, it is advised to use sequence-based classification methods that can allow FER systems to use the temporal information to recognize expressions from a set of frames.

Overall, the results showed that even the most accurate existing FER systems are not ready to face the challenges of a dynamic real-world. Thus future research in FER should focus on finding ways to handle the challenges highlighted in this research. It is hoped that the dataset collected in this study would become a useful benchmark for the evaluation of future FER systems.

Acknowledgments This research was supported by the MSIP, Korea, under the G-ITRC support program (IITP-2015-R6812-15-0001) supervised by the IITP, and by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2010-0020210).

References

1. Abidin Z, Alamsyah A (2015) Wavelet based approach for facial expression recognition. Int J Adv Intell Inf 1(1):7–14
2. Aleix M (1998) Martinez. The ar face database. CVC Technical Report
3. Bartlett MS, Littlewort G, Fasel I, Movellan JR (2003) Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: Conference on computer vision and pattern recognition workshop, 2003. CVPRW'03, vol 5, pp 53–53. IEEE
4. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 2, pp 568–573. IEEE
5. Bettadapura V (2012) Face expression recognition and analysis: the state of the art. arXiv:1203.6722
6. Bioid face db - humanscan ag, switzerland. https://www.bioid.com/About/BioID-Face-Database. Accessed: 2014-12-15
7. Chen L, Man H, Nefian AV (2005) Face recognition based on multi-class mapping of fisher scores. Pattern Recog 38(6):799–811

8. Dantcheva A, Chen C, Ross A (2012) Can facial cosmetics affect the matching accuracy of face recognition systems? In: 2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS), pp 391–398. IEEE
9. Face recognition and artificial vision group frav2d face database. http://www.frav.es/index.php/en/. Accessed: 2014-12-15
10. Fotosizer software. http://www.gomlab.com/eng/. Accessed: 2014-08-30
11. Fotosizer software. http://www.fotosizer.com/Download.aspx. Accessed: 2015-02-20
12. Garris MD (1994) Design, collection, and analysis of handwriting sample image databases. Encyclop Comput Sci Technol 31(16):189–213
13. Georghiades AS, Belhumeur PN, Kriegman D (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
14. Grgic M, Delac K, Grgic S (2011) Scface–surveillance cameras face database. Multi Tools Appl 51(3):863–879
15. Gross R, Matthews I, Cohn J, Kanade T, Baker S (2010) Multi-pie. Image Vis Comput 28(5):807–813
16. Happy S, Routray A (2015) Automatic facial expression recognition using features of salient facial patches. IEEE Trans Affect Comput 6(1):1–12
17. Happy SL, Routray A (2015) Robust facial expression classification using shape and appearance features. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), pp 1–5. IEEE
18. Jabid T, Md HK, Chae O (2010) Robust facial expression recognition based on local directional pattern. ETRI J 32(5):784–794
19. Jain V, Mukherjee A (2002) The indian face database. http://vis-www.cs.umass.edu/-vidit/IndianFaceDatabase
20. Kabir MdH, Jabid T, Chae O (2012) Local directional pattern variance (ldpv): a robust feature descriptor for facial expression recognition. Int Arab J Inf Technol 9(4):382–391
21. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Fourth IEEE international conference on automatic face and gesture recognition, 2000. Proceedings, pp 46–53. IEEE
22. Kang D, Han H, Anil K J, Lee S-W (2014) Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. Pattern Recogn 47(12):3750–3766
23. Kasinski A, Florek A, Schmidt A (2008) The put face database. Image Process Commun 13(3-4):59–64
24. Lisetti CL, LeRouge C (2004) Affective computing in tele-home health: design science possibilities in recognition of adoption and diffusion issues. In: Proceedings 37th IEEE Hawaii international conference on system sciences, Hawaii, USA
25. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 94–101. IEEE
26. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings, pp 200–205. IEEE
27. Marszalec E, Martinkauppi B, Soriano M, Pietika M et al (2000) Physics-based face database for color research. J Electron Imaging 9(1):32–38
28. Moore S, Bowden R (2009) The effects of pose on facial expression recognition. In: Proceedings of the British machine vision conference, pp 1–11
29. Nagaraja S, Prabhakar CJ (2014) Extraction of curvelet based rlbp features for representation of facial expression. In: 2014 international conference on contemporary computing and informatics (IC3I), pp 845–850. IEEE
30. Qi J, Gao X, He G, Luo Z, Yi W (2015) Multi-layer sparse representation for weighted lbp-patches based facial expression recognition. Sensors 15(3):6719–6739
31. Rivera AR, Castillo R, Chae O (2013) Local directional number pattern for face analysis: Face and expression recognition. IEEE Trans Image Process 22(5):1740–1752
32. Samaria FS, Harter AC (1994) Parameterisation of a stochastic model for human face identification. In: Proceedings of the Second IEEE Workshop on Applications of Computer Vision, 1994, pp 138–142. IEEE
33. Shbib R, Zhou S (2015) Facial expression analysis using active shape model. International Journal of Signal Processing, Image Processing and Pattern Recognition 8(1):9–22
34. Siddiqi MH, Ali R, Idris M, Khan AM, Kim ES, Whang MC, Lee S (2016) Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection. Multimedia Tools Appl 75(2):935–959

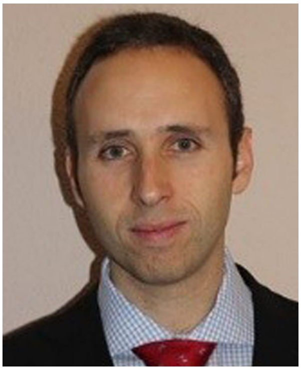
35. Siddiqi MH, Ali R, Khan AM, Kim ES, Kim GJ, Lee S (2015) Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimedia Systems 21(6):541–555
36. Siddiqi MH, Ali R, Khan AM, Park Y-T, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Process 24(4):1386–1398
37. Siddiqi MH, Lee S, Lee Y-K, Khan AM, Truc PTH (2013) Hierarchical recognition scheme for human facial expression recognition systems. Sensors 13(12):16682–16713
38. Sim T, Baker S, Bsat M (2003) The cmu pose, illumination, and expression database. IEEE Trans Pattern Anal Mach Intell 25(12):1615–1618
39. Singh R, Vatsa M, Bhatt HS, Bharadwaj S, Noore A, Nooreyezdan SS (2010) Plastic surgery: A new dimension to face recognition. IEEE Trans Inf Forensics Secur 5(3):441–448
40. Somanath G, Rohith MV, Vadana CK (2011) A dense dataset for facial image analysis. In: 2011 IEEE international conference on computer vision workshops (ICCV Workshops), pp 2175–2182. IEEE
41. Sung K-K, Poggio T (1998) Example-based learning for view-based human face detection. IEEE Trans Pattern Anal Mach Intell 20(1):39–51
42. The color feret database. http://www.nist.gov/itl/iad/ig/colorferet.cfm. Accessed: 2014-12-15
43. Thomaz CE, Giraldi GA (2010) A new ranking method for principal components analysis and its application to face image analysis. Image Vis Comput 28(6):902–913
44. Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, Chen F, Wang X (2010) A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimedia 12(7):682–691
45. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR), pp 529–534. IEEE
46. Wolf L, Hassner T, Taigman Y (2011) Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. IEEE Trans Pattern Anal Mach Intell 33(10):1978–1990
47. Wu X, Zhao J (2010) Curvelet feature extraction for face recognition and facial expression recognition. In: 2010 6th international conference on natural computation (ICNC), vol 3, pp 1212–1216. IEEE
48. Zhang B, Zhang L, Zhang D, Shen L (2010) Directional binary code with application to polyu near-infrared face database. Pattern Recogn Lett 31(14):2337–2344
49. Zhang L, Tjondronegoro D (2011) Facial expression recognition using facial movement features. IEEE Trans Affect Comput 2(4):219–229
50. Zhu Z, Ji Q (2006) Robust real-time face pose and facial expression recovery. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 681–688. IEEE

Muhammad Hameed Siddiqi is currently working as a Postdoctoral Research Scientist at the Networking Lab, Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea. He has completed his Bachelor of Computer Science (Hons) from Islamia College university of Peshawar, N-W.F.P, Pakistan in 2007, and Master and PhD from Ubiquitous Computing (UC) Lab, Department of Computer Engineering, Kyung Hee University, Suwon, Korea by 2012 and 2015, respectively. He was a Graduate Assistant at University Technology Petronas, Malaysia from 2008 to 2009. His research interest is Image Processing, Pattern Recognition, Machine Intelligence and Activity Recognition, Facial Expression Recognition.

Maqbool Ali received the B.S. degree with distinction in agricultural engineering from the University of Agriculture, Faisalabad, Pakistan, in 1999, the M.Sc. degree with distinction in computer science from the University of Agriculture, Faisalabad, Pakistan, in 2001, and the M.S. degree with distinction in Information Technology from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan, in 2013. From 2002 to 2005, he has served as an instructor in the field of computer science. From 2005 to 2014, he has been a researcher in the same field. Since March, 2014, he has joined Ubiquitous Computing Lab. to start his PhD at Dept. of Computer Engineering, Kyung Hee University, Republic of Korea. He is the author of 9 research papers. His research interests include Data Mining, Machine Learning, and Natural Language Processing.

Mohamed Elsayed Abdelrahman Eldib is currently Masters leading to PhD student at Biomedical Engineering department of Kyung Hee University, Korea. He has received his bachelor?s degree from Cairo University in the field of Biomedical Engineering in May 2011. His work as Biomedical Engineer at Al-Badr Engineering and Medical Company in 2012 was focused on Radiotherapy. He has also worked as trainee at Qasr Al-Aini hospital and Mostafa Kamal Military hospital in 2008 and 2009 respectively. His current research focuses on Medical Imaging Systems.

Asfandyar Khan received his Ph.D. degree from the Department of Computer and Information Sciences, University Technology Petronas, Malaysia in 2011. He is now working as a faculty member with the Department of Computer Science, University of Science & Technology Bannu, Pakistan. His research interest includes Wireless Sensor Network, Bioinformatics, Pattern Recognition.



Oresti Banos (MSc in Telecommunications Engineering in 2009, MSc in Computer Network Engineering in 2010, MSc in Electrical Engineering in 2011, PhD in Computer Science in 2014, all with honors from the University of Granada, Granada, Spain) is currently working as a Postdoctoral Research Scientist at the Ubiquitous Computing Lab (UCLab), Kyung Hee University, Korea. He has a wide range of experience in wearable, ubiquitous, pervasive and mobile computing with a particular focus on digital health and wellness applications. His area of expertise lies within pattern recognition and machine learning for probabilistic modeling of human behavior and context-awareness. He has a strong background on multi-modal sensor data processing, data fusion and transfer learning. He is especially interested in the field of robust, adaptive, opportunistic and intelligent digital health systems. He is author of more than 35 papers, most of them indexed in top-ranked international conferences and journals. He has also super-vised several MSc projects in health, wellbeing and biomedical domains. He serves as reviewer and program committee member in several international journals and conferences, including Medical Engineering and Physics (Elsevier), Journal of Medical and Biological Engineering (Springer), Pervasive and Mobile Computing (Elsevier), Sensors (MDPI), UBICOMP, ISWC and IWBBIO.

Adil Mehmood Khan received his Ph.D. degree from the Department of Computer Engineering of Kyung Hee University, Republic of Korea in 2011. He is now working as a faculty member with the Department of Computer Science, Innopolis University, Kazan, Russia. His research interest includes Pattern Recognition, Signal Processing, Ubiquitous Computing, and Machine Learning.

Sungyoung Lee received his B.S. from Korea University, Seoul, Korea. He got his M.S. and Ph.D. degrees in Computer Science from Illinois Institute of Technology (IIT), Chicago, USA in 1987 and 1991 respectively. He has been a professor in the department of Computer Engineering, Kyung Hee University, Korea since 1993. He is a founding director of the Ubiquitous Computing Laboratory, and has been affiliated with a director of Neo Medical ubiquitous-Life Care Information Technology Research Center, Kyung Hee University since 2006. Before joining Kyung Hee University, he was an assistant professor in the Department of Computer Science, Governors State University, llinois, USA from 1992 to 1993. His current research focuses on Ubiquitous Computing and Applications, Wireless Ad-hoc and Sensor Networks, Context-aware Middleware, Sensor Oper-ating Systems, Real-Time Systems and Embedded Systems, Activity and Emotion Recognition. He is a member of ACM and IEEE.

Hyunseung Choo received the B.S. degree in mathematics from Sungkyunkwan University, Korea in 1988, the M.S. degree in computer science from the University of Texas at Dallas, USA in 1990, and the Ph.D. degree in computer science from the University of Texas at Arlington, USA in 1996. From 1997 to 1998, he was a patent examiner at Korean Industrial Property Office. Since 1998, he has joined the College of Information and Communication Engineering, Sungkyunkwan University, and is an associate professor and director of Convergence Research Institute. Since 2005, he is director of Intelligent HCI Convergence Research Center (eight-year research program) supported by the Ministry of Knowledge Economy (Korea) under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment. His research interests include wired/wireless/optical embedded networking, mobile computing, and grid computing. He is vice president of Korean Society for Internet Information (KSII). He has been editor-in-chief of the Journal of KSII for three years and journal editors of Journal of Communications and Networks, ACM Transactions on Internet Technology, International Journal of Mobile Communication, Springer Verlag Transactions on Computational Science Journal, and editor of KSII Transactions on Internet and Information Systems since 2006. He has published over 200 papers in international journals and refereed conferences. He is a member of IEEE and ACM.