This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

Efficient Image Recognition and Retrieval on IoT-Assisted Energy-Constrained Platforms from Big Data Repositories

Irfan Mehmood, Member, IEEE, Amin Ullah, Student Member, IEEE, Khan Muhammad, Member, IEEE, Der-Jiunn Deng, Weizhi Meng, Fadi Al-Turjman, Member, IEEE, Muhammad Sajjad, Victor Hugo C. de Albuquerque, Member, IEEE

***Abstract*— The advanced computational capabilities of many resource constrained devices such as smartphones have enabled various research areas including image retrieval from big data repositories for numerous IoT applications. The major challenges for image retrieval using smartphones in an IoT environment are the computational complexity and storage. To deal with big data in IoT environment for image retrieval, this paper proposes a light-weighted deep learning based system for energy-constrained devices. The system first detects and crops face regions from an image using Viola-Jones algorithm with additional face and non-face classifier to eliminate the miss-detection problem. Secondly, the system uses convolutional layers of a cost effective pre-trained CNN model with defined features to represent faces. Next, features of the big data repository are indexed to achieve a faster matching process for real-time retrieval. Finally, Euclidean distance is used to find similarity between query and repository images. For experimental evaluation, we created a local facial images dataset, including both single and group facial images. This dataset can be used by other researchers as a benchmark for comparison with other real-time facial image retrieval systems. The experimental results show that our proposed system outperforms other state-of-the-art feature extraction methods in terms of efficiency and retrieval for IoT-assisted energy-constrained platforms.**

***Index Terms*— Image Retrieval, Internet of Things (IoT), Big Data, Convolutional Neural Network, Energy-Constrained Platforms, Deep Learning**

I. INTRODUCTION

In today's modern era of technology, dealing with multimedia data has become a very burning issue of research for which many scientists tried to contribute in computer vision and IoT society. Usage of smartphones and other IoT assisted devices are increasing rapidly results in expanding the collection rate of images exponentially. Thus, efficient mechanisms for managing, searching, retrieving and indexing of image big data repositories are needed [1, 2]. Traditional approaches of annotating images with text for indexing and retrieval have a lower accuracy rate. Moreover, manual labeling of image big data is tiresome and fails to express the exact contents of an image because more than one objects can be referred by the same words such as car, truck, and bus can be labeled as vehicle [3]. Therefore, content-based image retrieval (CBIR) [4, 5] has gained considerable attention of researchers from the past decade. Many CBIR methods have been developed in the field of education, entertainment, agriculture, defense, IoT-assisted surveillance, and medical sciences [6, 7]. CBIR indexes images using features extracted from its color, texture, shape, and spatial layout, which are known as low-level features. Low-level features are unable to represent the entire semantics of an image, because two different images may have same low-level features, but they are easy to compute and implement. Features extraction from an image which compactly describes its content is one of the challenging task in CBIR for which many image representation techniques have been developed [8-11]. Histogram of oriented gradients (HOG) [9] calculates orientation and magnitude of pixels for localized portions. Scale-invariant features transform (SIFT) [10] and speed up robust features (SURF) [11] sought interest points in the image and localize features for those key points to represent an image. In current research arena of features extraction, images and videos are mostly represented by CNN features [12-15], due to immense increase in image classification accuracy using CNNs for ImageNet [16]. ImageNet is a gigantic dataset contains more than one million images with one thousand classes. CNN has the ability to find the hidden patterns in an image with assistance of its millions of parameters (weights and biases) to effectively represent an image. However, traditional CNN models contain millions of parameters that require high computation to extract features, reducing their suitability for energy-constrained devices. Therefore, researchers are trying to implement efficient and robust CNN models on various resource-constrained platforms such as raspberry pi and smartphones for IoT applications [17-19].

The IoT technology for connecting smart devices such as smartphones and various sensors is emerging rapidly. A large amount of data is generated from these smart devices such as vision sensors that can be used for various applications

Manuscript received November 1, 2018; Accepted: XXX, Published: XXXX. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (Ministry of Science and ICT) (No. 2018R1C1B5086294). This paper was recommended by Associate Editor XYZ. (*Corresponding author: Muhammad Sajjad*)

Irfan Mehmood is with Department of Software, School of Electronics and Information Engineering, Sejong University (Email: irfanmehmood@ieee.org)

Amin Ullah and Khan Muhammad are with Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Republic of Korea (Email: aminullah@ieee.org, khan.muhammad@ieee.org)

Der-Jiunn Deng is with the Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua 500, Taiwan (e-mail: djdeng@cc.ncue.edu.tw)

Weizhi Meng is with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark. (Email: weme@dtu.dk)

Fadi Al-Turjman is with the College of Engineering, Antalya Bilim University, 07190 Antalya, Turkey (Email: fadi.alturjman@antalya.edu.tr)

Muhammad Sajjad is with Digital Image Processing Laboratory, Department of Computer Science, Islamia College, Peshawar, Pakistan (Email: muhammad.sajjad@icp.edu.pk)

Victor Hugo C. de Albuquerque is with Graduate Program in Applied Informatics at the Universidade de Fortaleza, Fortaleza/CE, Brazil (Email: victor.albuquerque@unifor.br)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

including security, privacy, person re-identification, and image retrieval. The smartphones connected to IoT technology assists people in many daily life problems, but it has very limited storage capabilities and computational power. The smartphone users cannot execute any complex programs, particularly while dealing with high resolution images. Therefore, most of the IoT connected smart devices transmit data to cloud servers for efficient processing and its storage. Thus, the control of image data transfers from local smartphones or other devices to cloud servers. Currently, a large number of research institutes are working on image retrieval systems for the benefit of the society. For instance, a secure image retrieval system in IoT is proposed by [20], where they used resource-constrained clients to move the preprocessing of images to cloud. On cloud server, the image search is performed, thus helps in reducing the cost for the end user. Another encrypted image retrieval system in IoT with multi-user authentication is proposed by [21]. This system is lightweight, enabling content-based search through decrypted images. They represented images using local visual features, followed by Euclidean distance to measure the similarity between two feature vectors for retrieval. Rahim et al. [22] encrypted images using light-weight secure encryption algorithm on smartphone prior to sending it towards cloud for binary compact codes generation for image retrieval. In addition to vision sensor data, other smartphone sensors such as gyroscope, accelerometer, and fingerprint sensors are also integrated together in an IoT platform to serve humanity by performing various tasks.

Audio calls and messaging through cell phones, used a decade ago, have now transformed into live video chats with other features like location tracing through global position system (GPS) [23]. These smartphone features enable various IoT applications such as homes intelligence, IoT in healthcare, IoT in vehicles, smart grids and a lot more [24]. The extensive features of a smartphone like high-resolution camera, long battery life, and huge storage enables a user to capture a massive number of images and record videos. Researchers from different areas are utilizing smartphone's resources such as an accelerometer, gyroscope, camera, and fingerprint sensors for different tasks [25]. Zualkernan et al. [26] collected data from accelerometer readings, analyzed several features of typing actions such as quickness and delay between typing strokes for training a classifier for human emotions prediction. Chetty et al. [27] proposed a technique for human activity recognition using the inertial sensors. They used the concept of information theory-based features ranking algorithm. A precise scalable mobile image retrieval technique is presented by Yang et al [28] which retrieves images in two steps. Firstly, it determines the relevant images based on visual similarity and then it acquires scalable retrieval by subtracting contextual saliency from retrieved images. Jonathon et al. [29] proposed a smartphone based CBIR and object recognition technique and claimed that the method can work on degraded images such as noise affected and different transformations. Their technique extracts SIFT features from salient regions of images that are indexed through the vector-space model with a two-stage ranking technique for efficient retrieval of images.

In generic CBIR concept, images are retrieved according to the user desired query of certain category [30, 31] from large scale image databases. A recent work [3] developed a CBIR system in which images are represented by fusing salient color features with rotation invariant texture features (ISC&RIT). In this method, color features are extracted using HSV color quantization histogram and texture features using rotated local binary pattern (RLBP) [32]. Due to the robust representation of an image by color and texture, they used it for general category images retrieval. These techniques extract global image features of general image category and are not appropriate to represent local facial features. Moreover, these techniques are based on handcrafted features, which are not able to capture full semantics of image and are computationally expensive.

In this paper, we retrieve images based on user provided query face. For instance, a user can retrieve his images from the IoT-assisted smartphone gallery where he is alone or in a group with friends and family members. The proposed system detects face using Viola-Jones [33] technique, crops it from the image, inquire for true positive detection and extracts convolutional features of an efficient pre-trained CNN model. A similar process is applied for all the photographs present in the big data repository. To retrieve the same face images, we calculate Euclidean distance between the query and database face features. The accurate detection of faces in the proposed system is because of scrutiny for false and true positive using suggested face and non-face trained classifier. The local features extracted from convolutional layers of proposed CNN proves to be the best representative of different parts of face. The proposed CNN model is designed to be cost-effective for energy-constrained devices for IoT applications. Thus, the above properties allow this system to detect the faces with high accuracy and represent faces effectively using convolutional features.

The paper is organized as follows: The proposed IoT assisted image retrieval framework is explained in Section II. Experimental results and evaluation of our technique with state-of-the-art techniques are discussed in Section III. Section IV concludes the paper with future directions.

II. The Proposed Framework

This section illustrates the concept of retrieving images based on convolutional facial features for IoT assisted energy-constrained devices. Also, the framework of features extraction using intermediate layers of cost-effective pre-trained CNN is described in detail. The proposed system has three core steps. Firstly, faces $\mathcal{F}_N$ are detected from image $\mathcal{I}$ using Viola-Jones algorithm and cropped. Each image may contain different number of faces. Therefore, we have indexed each cropped face $\mathcal{C}_\mathcal{F}$ with its associated image $\mathcal{I}$. Secondly, the cropped faces $\mathcal{C}_\mathcal{F}$ are fed to the CNN model. The two main layers i.e., Conv4 and Pool3 of CNN are utilized for feature extraction. Convolution features ω^{Conv4} and ω^{Pool3} are fused together as $\mathcal{T}_\mathcal{F}$ for face representation. These two steps are repeated for whole dataset images and for each image the features of faces are stored in features database $\Theta_\mathcal{F}$. Finally, Euclidean distance is utilized to measure the similarity score between query face features $\mathcal{F}_q$ and face features database $\Theta_\mathcal{F}$ in real-time. Framework of the proposed system is shown in Fig. 1. The input/output parameters of the proposed system in addition to abbreviations used in this paper are given in Table I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

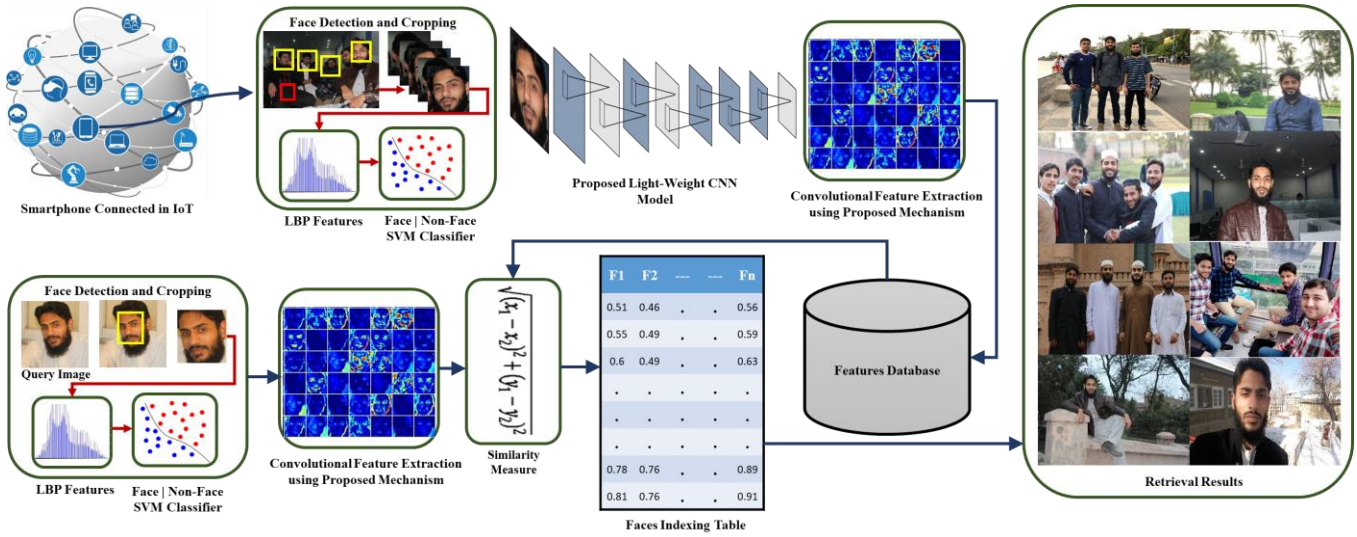


Fig. 1. The framework of the proposed IoT assisted facial-features based image retrieval system where detected faces through Viola-Jones algorithm are cropped and fed to face/non-face SVM classifier to eradicate the miss-detection problem of Viola-Jones algorithm. Next, the cropped faces are input to the proposed light-weight CNN model to extract convolutional features. Finally, the extracted features are matched with big data repository features via Euclidean distance to retrieve similar faces.

Table I.
Model parameters and abbreviations used throughout the paper.

Description of model parameters:	
F_q	Query face
I	Image in process
ω^{Conv4}	Convolution layer four features
T_F	Fused features
$\times_I$	Faces indexing table
S2	Sum of pooling feature maps
F_N	Total faces in image
C_F	Cropped face from *I*
ω^{Pool3}	Pooling layer three features
Θ_F	Features database
S1	Sum of convolutional feature maps
FV	Final features vector

A. *Preparation and Face Detection*

The proposed system is totally based on the face regions in the image. Therefore, face detection is the first challenging step, as images are mostly taken with different illumination changes, scenes, poses, and viewpoints on a smartphone. We have used a well-known Viola-Jones [33] face detection algorithm for face detection. The reason for using Viola-Jones is that it is open source and fast face detection algorithm which helps us to maintain low complexity on a smartphone in IoT network. However, it has the limitation of false-positives in complex background images shown in Fig. 2. Therefore, to overcome the effect of false-positive detection in our system we have trained a two class (Face, Non-Face) classifier as a verification step for face detection. It helps to analyze only true-positive detected faces in the image.

A binary class SVM is trained on face and non-Face images. Faces are cropped using Viola-Jones method while non-face images are collected from different non-face regions of images. A dataset of five hundred face images and five hundred non-face images are prepared for training a classification model. Local binary patterns (LBP) texture features are extracted for training a binary classifier which can discriminate between true-positive and false-positive detection of faces. The prepared data is trained and tested on three algorithms including linear SVM, quadratic SVM, and decision Tree with ten folds cross-validation. We got 92%, 95%, and 90% validation accuracy from these classifiers, respectively. Finally, in the proposed system we have used quadratic SVM classifier which has less false-positive score and higher true-positives. Fig. 2 (a) shows the average face and non-face features representation. Which are classified using trained quadratic SVM classifier and visualized in Fig. 2 (b).

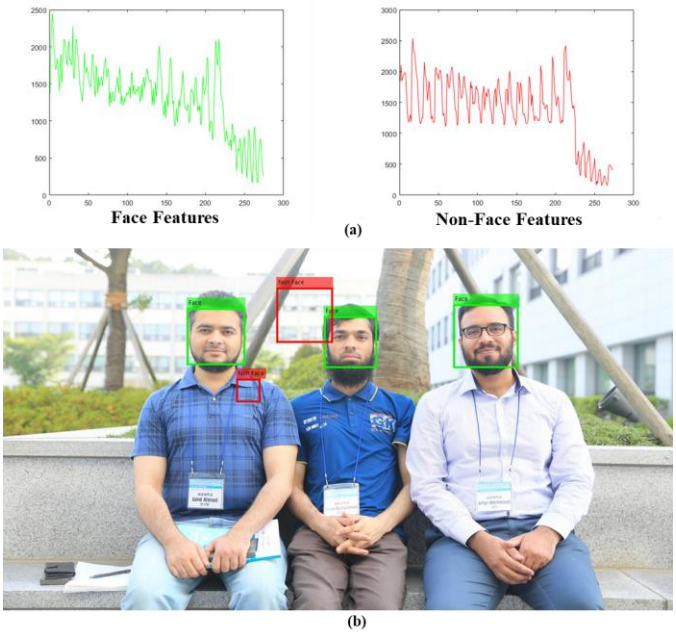


Fig. 2. Result of the suggested face, non-face classification module.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

B. Proposed facial features extraction

In recent years, CNN based approaches retain the overwhelming benefits compared to some previous handcrafted features extraction methods, particularly in semantics representation of image[12, 34, 35]. In the deep learning based approaches, the features of the earlier layers contain higher spatial resolution for precise local features, while the features in later layers indicates more semantic or global information [36]. The activations that are the output of CNN layers are interpreted as visual features. We have utilized local features of convolution layer four and pooling layer three for more generic face representation [37]. The architecture of pre-trained CNN is given in Fig. 3. It shows three new layers after final pooling S1, S2, and FV where S1 layer is the sum of 15×15 conv4 feature maps with 384 channels. S2 is the sum of 7×7 pool3 feature maps with 256 channels. These two layers are converted to one dimensional 225 and 49 feature vector, respectively. The weights size, strides, padding, channels, and outputs of the network is given in Table. II. The proposed system uses pre-trained CNN model which is trained on VGG face dataset [38]. The dataset has 2597 subjects of different TV and movie actors having more than 0.8 million images. Each subject has images ranging from 120 to 250. We have trained a CNN model on VGG face dataset having similar architecture as AlexNet [39] CNN model. We have modified the size of the input image from 227×227 to 128×128 as the detected face in an image has lower resolution. Further, we have changed the size of convolutional kernels from 11×11 to 5×5 because the small size filter can learn more tiny discriminative changes in the visual data [40] which is more suitable for face images. The reason for training a new model on VGG face dataset instead of using a pre-trained AlexNet model is that the original AlexNet model is trained on visual data of general categories which is able to extract discriminative features from data of diverse categories. However, visual information in faces is very similar for everyone. Therefore, the general category CNN models are not able to extract discriminative visual features for facial data. Thus, we first trained a CNN model on face image dataset from scratch and then we performed our proposed mechanism for convolutional feature extraction to represent face images. In the proposed system, we have extracted convolutional features and skipped the fully connected layers and Softmax layer. The reason for eliminating these layers is that fully connected layers represent more global features of the image. Where we need local information of face part such as eyes, nose, and lips etc. which can be represented more accurately using the local features. Convolutional features are capable of extracting local features [40]. The effect of convolutional features on face can be seen in Fig. 4. Secondly, as the image retrieval is not a classification problem, therefore, we have eliminated Softmax layer. The architecture of proposed technique is numerically described in Table. II. The proposed CNN has five convolutional layers and three pooling layers. The kernels size, strides, and padding are changed from original AlexNet [39] model because the image we feed to CNN is 128x128 instead of 227x227. The sizes of feature maps are different but the number of channels are same.

Fig. 3 shows three new layers after final pooling S1, S2, and FV. Where S1 layer is the sum of 15×15 conv4 feature maps with 384 channels. S2 is the sum of 7×7 pool3 feature maps with 256 channels. These two layers are converted to one dimensional 225 and 49 feature vector respectively.

Table II.
Weights and outputs of the proposed CNN model used for convolutional features extraction.

Layers	Conv1	Pool1	Conv2	Pool2	Conv3	Con4	Con5	Pool3	S1	S2	FV
Kernel	5×5	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
Stride	2	2	1	2	1	1	1	2	-	-	-
Pad	0	0	2	0	1	1	1	0	-	-	-
Channel	96	96	256	256	384	384	256	256	1	1	1
Output	62×62×96	31×31×96	31×31×256	15×15×256	15×15×384	15×15×384	15×15×256	7×7×256	225	49	274

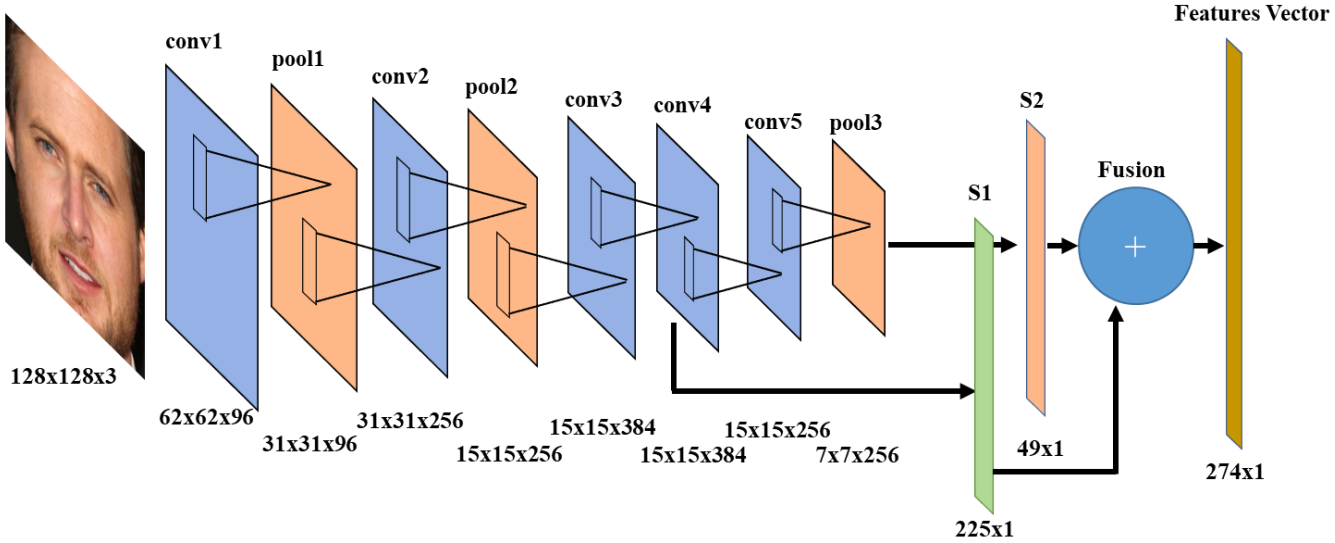


Fig 3: Proposed architecture for facial feature extraction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

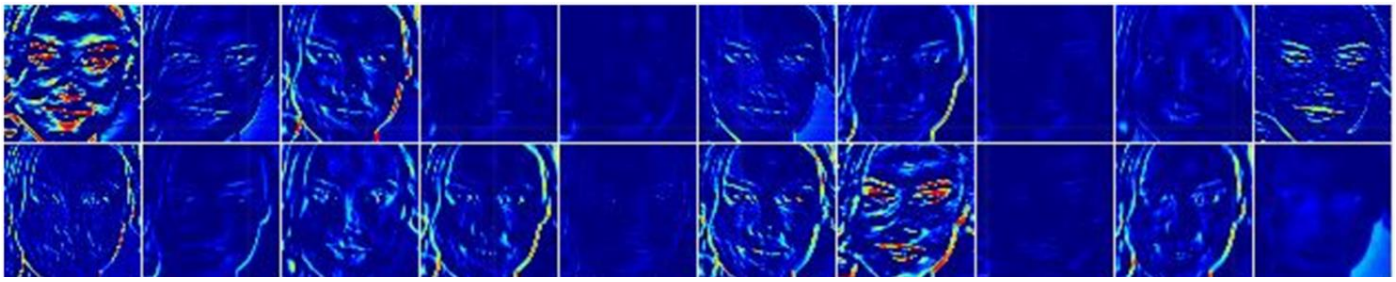


Fig 4: Face representation through convolution layer four feature maps.

Algorithm: Image retrieval and recognition for IoT assisted energy-constrained devices.

Input: Query Face $\mathcal{F}_q$, big data repository Φ

1. Select one image $\mathcal{I}$ from big data repository Φ.
2. Detect faces $\mathcal{F}_N$ in the image $\mathcal{I}$.
3. Crop faces C_F from $\mathcal{F}_N$.
4. Feed C_F to proposed CNN model.
5. Extract LBP features from C_F
6. Feed LBP to train SVM model for face and non-face classification.
   If prediction == face
   Extract FV from covn4 ω^{Conv4} and pool3 ω^{Pool3}
   Save FV to features database Θ_F
   Else
   Do nothing
   end
7. Repeat step 1 to 6 for all images of the big data repository Φ.
8. Select query face image $\mathcal{F}_q$.
9. Apply step 2 to 5 on $\mathcal{F}_q$.
10. Calculate the Euclidean distance between $\mathcal{F}_q$ and Θ_F.
11. Update faces indexing table $\times_I$

Output: Retrieve the faces having smaller Euclidean distance with $\mathcal{F}_q$ in $\times_I$.

Finally, S1 and S2 are fused to make features vector for face representation. To overcome the complexity of system the proposed technique extract features once from all images of the big data repository and make a feature database ΘF. The similarity between query face and faces in the repository is calculated using Euclidean distance of the features. The distance values are stored in faces indexing table $\times_I$ and faces with the minimum distance to the query face are retrieved as a final output of the system.

III. Experimental Results and Discussion

In this section, the proposed approach has been evaluated using several image retrieval metrics including precision, recall and mean average precision (MAP). We performed several experiments using actual smartphone device as an emulator i.e., LG-G4, having Android 6.0 marshmallow, software version H81120x installed on it. The processing components of the device include snapdragon 808 hexa-core processor with 3GB of RAM and 16 MP rear camera and 8 MP selfie camera. We have arranged our own dataset for the evaluation of the proposed system. Because there is no available dataset containing group images of individuals, friends, and family. We collected nine hundred images of our laboratory members taken on different occasions. It contains twenty common subjects, where for each subject we have individual and group photos. It is more challenging because the images are taken with different illumination changes, indoor, and outdoor environments. The proposed system is evaluated using precision, recall score [41] and MAP score [42]. Our system is also assessed and compared with different handcrafted feature based methods including the local binary pattern (LBP) [43], HOG [44], SURF [11], and color texture fused features [45] based techniques.

$$\mathcal{P} = \frac{relevent\_images\_retrieved\_(\mathcal{R}_I)}{Number\_of\_selected\_images\_from\_gallery\_(S_I)} \quad (1)$$

$$\mathcal{R} = \frac{relvent\_images\_retrived\_(\mathcal{R}_I)}{overall\_relevent\_images\_in\_gallery\_(\mathcal{N}_I)} \quad (2)$$

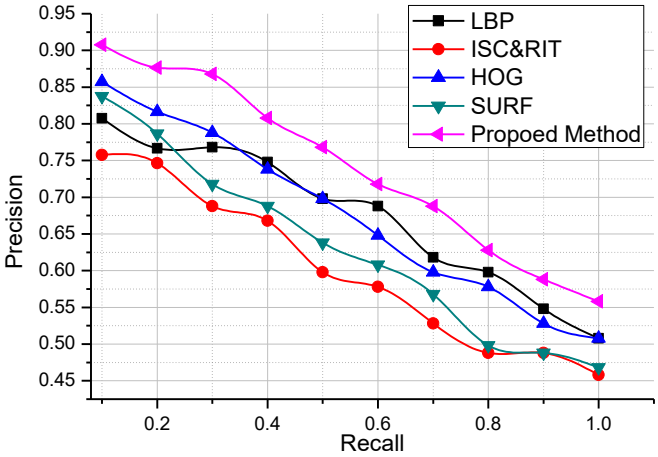


Fig. 5: Precision and recall graph of the proposed system with comparison against well-known features extraction methods.

A well-known information retrieval evaluation method "precision and recall" is used for the performance assessment of the proposed system. The precision $\mathcal{P}$ is the ratio between the amount of the relevant images $\mathcal{R}_I$ retrieved and the number of selected retrieved images S_I from big data repository. It calculates the positive predictive values of the information retrieval system. Recall $\mathcal{R}$ is defined as the ratio between the amount of the relevant images $\mathcal{R}_I$ retrieved and the total number of the relevant images $\mathcal{N}_I$ from repository. Both precision and recall are therefore based on an understanding and measure of relevance. $\mathcal{P}$ and $\mathcal{R}$ are computed using Eq. 1 and Eq. 2 [46].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

We have conducted several experiments using different handcrafted features before the evaluation of deep features and convolutional features. As handcrafted features are easy to compute with fewer execution resources and time. However, its performance is not effective for image representation. Fig. 5 shows the performance of different features extraction techniques. Patterns finding techniques in face images such as LBP, HOG, SURF, and SIFT that are already used in many face recognition and facial expression analysis methods are not working effectively for the proposed system. Firstly, we have evaluated our previous work (ISC&RIT) [3] for the proposed idea. ISC&RIT combines salient color and rotation invariant texture feature for image representation, but it has low accuracy for the proposed system. HOG features achieved second highest accuracy overlapping on same recall level with LBP features. SURF, color, and texture fused features-based methods are well behind the proposed technique. It can be seen from Fig. 5 that for low recall level every feature extraction method has a high precision score. But as the recall level increases it indicates that the number of selected images from the repository is also increasing, so the precision score decreases gradually. This is because when we select less number of images for calculation precision score, it means that we are selecting the higher ratio similar images from the repository. Therefore, the accuracy is high, when we are increasing the number of selected images from repository, so the similarity score between query face image and features database is very low. The proposed method has higher accuracy throughout the recall level as compared to handcrafted features extraction methods as well as deep features of pre-trained CNN models.

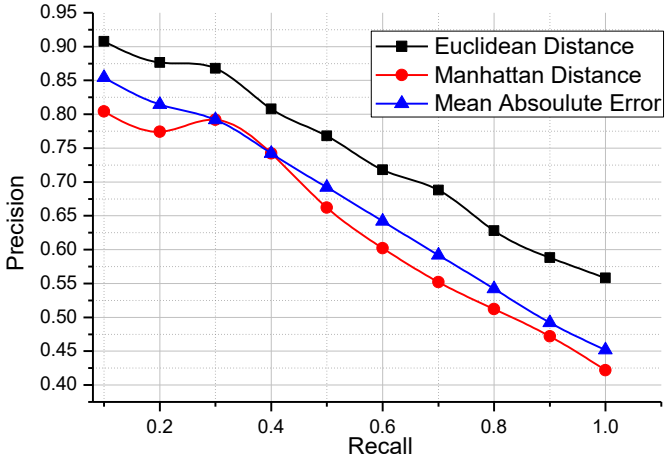


Fig. 6: Precision values of various similarity measuring metrics on the proposed system for different recall levels.

The retrieval performance of proposed technique is also assessed using different similarity and dissimilarity matrices such as Euclidean distance, Mean Absolute Error, and Manhattan distance. After analyzing results of these three metrics we have used Euclidean distance because of its high precision score. Fig. 6 shows retrieval performance of three distance metrics, where the precision scores of Manhattan distance and Mean Absolute Error are closed to each other's and overlaps on some recall levels. The Euclidean distance reached high precision score all over the recall axis and beats the other methods with a higher margin. This is because the Euclidean distance is efficient in calculating pairwise matching of features. Therefore, we have used Euclidean distance for comparison between query face features and database features throughout the experiments.

Deep learning and CNN has excessive ability to represent an image semantically. Therefore, we have also evaluated deep features of pre-trained CNN model including fully connected layers FC7 and FC8 for the compassion with proposed convolutional features for face-based image retrieval problem. Fig. 7 shows the performance compassion between deep features and convolutional features. On low recall level, different method performs well and get the high precision score. However, as deep features of CNN represent more global features of a given image, and proposed problem deal with faces. Therefore, when we extract global information from faces, so we get almost same features representation for every face. On the other hand, convolutional features represent local features of a given image. Consequently, we can get information about the face, eyes, nose, mouth and other parts in feature representation of the face. Thus, from Fig. 7 we can see that both fully connected layers overlap each other throughout the graph. Convolutional features improved overall precision score on each recall level. Due to this reason, we have used convolutional features for face representation for the proposed problem.

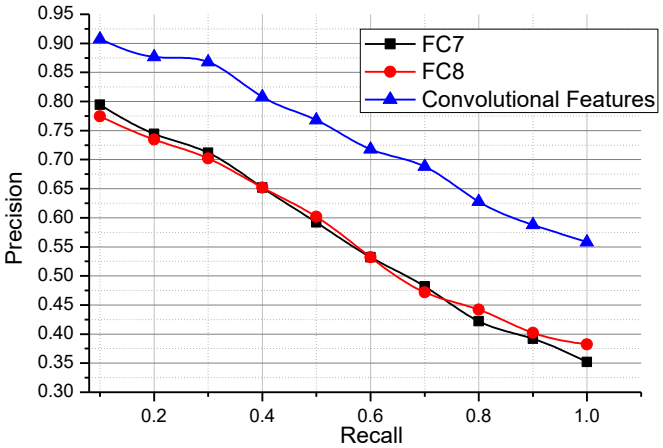


Fig. 7: Performance comparison of deep features against convolutional features for face-based image retrieval problem.

The MAP is another most common used metric for the evaluation of retrieval systems. MAP value is calculated as follows [30]:

$$MAP = 1/L\sum_{j=1}^{L} AP_j \quad (3)$$

Where L represents the total number of queries for evaluation of the system and AP is the mean precision, representing mean precision values of the similar results for all queries. It is computed as follows:

$$AP = 1/S\sum_{j=1}^{S} P_j \quad (4)$$

The P_j precision value for all the similar images S. The MAP score ranks the retrieval system and it ranges from 0 to 1. The MAP score near to 1 means system is of higher rank and near to zero means the retrieval is of lower rank. Table III. shows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

percentage MAP scores of the proposed technique and other state-of-the-art methods. The MAP scores are calculated up to five recall levels on 40 different face query images from the dataset. The proposed method achieved a maximum score of 85.39 on the local dataset. On the other hand, deep features FC7 with 79.74 and FC8 with 80.63 MAP scores have better performance from handcrafted feature extraction methods.

Table III.
Comparison of the proposed system with other state-of-the-art techniques using MAP score.

Methods	**MAP (%)**
LBP	72.51
ISC&RITP	67.81
HOG	77.32
SURF	70.14
FC7 Features	79.74
FC8 Features	80.63
Convolutional Features	**85.39**

Fig. 8 and Fig. 9 show retrieval results of two face queries. In both figures, the first image is the query image while other images are the retrieved similar images from the database. The Euclidean distance score between query face and features database of faces are given on the top of each image. Lower distance indicates higher similarity to the query image and vice versa. In Fig. 8 and Fig. 9 we can see that proposed method retrieved mostly similar images. However, there are some images which are not related to query face, but they are retrieved because their visual content such as texture and shape are similar to the query face. Fig. 8 shows positive side and Fig. 9 represent more challenging part of the proposed technique. One of the major problems in face image representation and recognition is the beard men. Therefore, in Fig.8 where the query face is clean shaven which make it easy for the feature extraction method for efficient representation of face. While on the other hand, we have several people in repository images who have a beard and at the time of face detection and face representation we get false-positive detection for those faces. Therefore Fig. 9 has false retrieval of some images. In future work we will try to overcome the problem of face detection in challenging scenarios, also fine-tune the proposed model such that it can efficiently represent both clean shaven and beard men.

Fig. 8: Retrieval results of the proposed system using convolutional features.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2896151, IEEE Internet of Things Journal

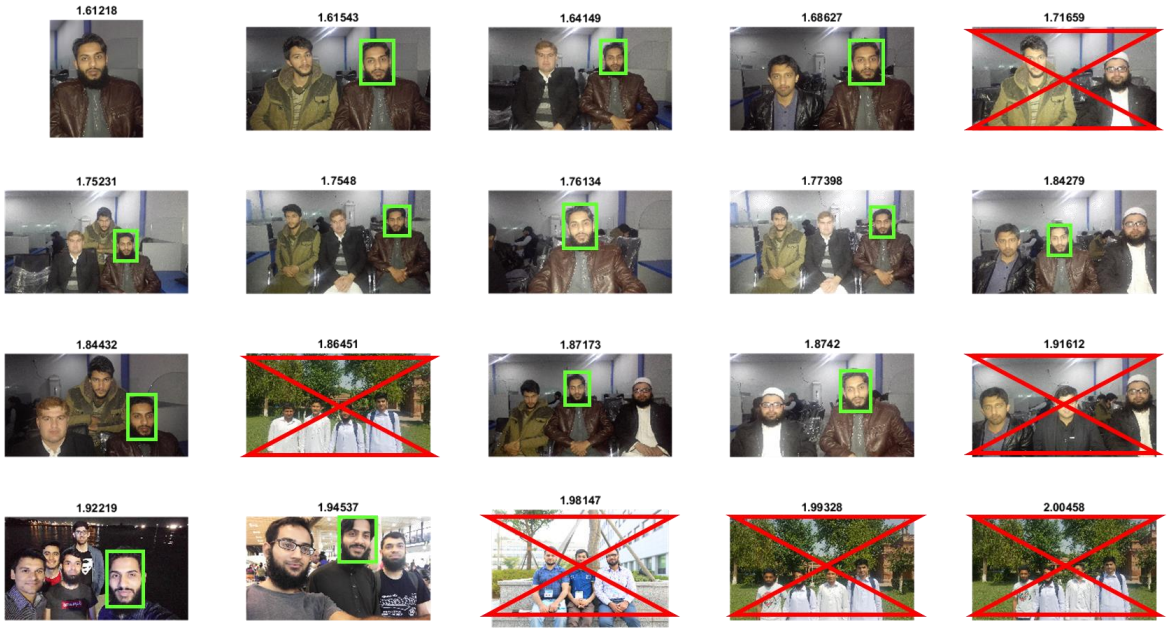


Fig. 9: Retrieval results of the proposed system using convolutional features.

IV. Conclusion and Future Work

In this paper, we presented a real-time face query-based image retrieval system for IoT-assisted energy-constrained devices. The proposed technique analyzes the face regions of all images containing group and individual photos. The face detection algorithm we used had a problem of false-positive detection, which we solved using binary classifier for analyzing only true-positive faces. Secondly, we used an efficient CNN for feature extraction where convolutional layer 4 and pooling layer 3 are utilized for face image representation. We have conducted various experiments for image representation and similarity measure. For image representation, we have analyzed features of convolutional and fully connected layers. For similarity measures, as compared with MSE and Manhattan distance, Euclidean distance could provide higher accuracy on our proposed convolutional features extraction mechanism. The proposed method is very effective in terms of both accuracy and complexity, which can be a part of IoT assisted energy-constrained devices [47] for efficient real-time image retrieval system. In future, we plan to analyze hash-based image representation techniques [48-50] which will help us in storing features on small capacity devices and achieve more robust retrieval performance.

References

[1] S. Rho, "Efficient Object-Based Distributed Image Search in Wireless Visual Sensor Networks," *JOURNAL OF PLATFORM TECHNOLOGY*, vol. 5, pp. 27-39, 2017.

[2] K. Muhammad, R. Hamza, J. Ahmad, J. Lloret, H. H. G. Wang, and S. W. Baik, "Secure surveillance framework for IoT systems using probabilistic image encryption," *IEEE Transactions on Industrial Informatics*, 2018.

[3] M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, and S. W. Baik, "Integrating salient colors with rotational invariant texture features for image representation in retrieval systems," *Multimedia Tools and Applications*, vol. 77, pp. 4769-4789, 2018.

[4] M. Tzelepi and A. Tefas, "Deep convolutional learning for Content Based Image Retrieval," *Neurocomputing*, vol. 275, pp. 2467-2478, 2018.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal*, vol. 1, pp. 22-32, 2014.

[6] X. Chang, Z. Ma, M. Lin, Y. Yang, and A. G. Hauptmann, "Feature interaction augmented sparse learning for fast kinect motion detection," *IEEE Transactions on Image Processing*, vol. 26, pp. 3911-3920, 2017.

[7] Z. A. Abduljabbar, H. Jin, A. Ibrahim, Z. A. Hussien, M. A. Hussain, S. H. Abbdal, *et al.*, "Privacy-preserving image retrieval in IoT-cloud," in *Trustcom/BigDataSE/I SPA, 2016 IEEE*, 2016, pp. 799-806.

[8] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE transactions on neural networks and learning systems*, vol. 28, pp. 2294-2305, 2017.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91-110, 2004.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404-417, 2006.

[12] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155-1166, 2018 2017.

[13] J. Ahmad, K. Muhammad, S. Bakshi, and S. W. Baik, "Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets," *Future Generation Computer Systems*, vol. 81, pp. 314-330, 2018.

[14] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE transactions on cybernetics*, vol. 47, pp. 1180-1197, 2017.

[15] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, pp. 1617-1632, 2017.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255.

[17] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for

effective disaster management," *Neurocomputing*, vol. 288, pp. 30-42, 2018.

[18] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional Neural Networks Based Fire Detection in Surveillance Videos," *IEEE Access*, vol. 6, pp. 18174-18183, 2018.

[19] K. Muhammad, J. Ahmad, Z. Lv, and P. Bellavista, "Efficient Deep CNN-Based Fire Detection and Localization in Video Surveillance Applications."

[20] H. Yan, Z. Chen, and C. Jia, "SSIR: Secure similarity image retrieval in IoT," *Information Sciences*, vol. 479, pp. 153-163, 2019/04/01/ 2019.

[21] M. A. Al Sibahee, S. Lu, Z. A. Abduljabbar, A. Ibrahim, Z. A. Hussien, K. A.-A. Mutlaq, *et al.*, "Efficient encrypted image retrieval in IoT-cloud with multi-user authentication," *International Journal of Distributed Sensor Networks*, vol. 14, p. 1550147718761814, 2018.

[22] N. Rahim, J. Ahmad, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Privacy-Preserving Image Retrieval for Mobile Devices with Deep Features on the Cloud," *Computer Communications*, 2018.

[23] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, pp. 450-465, 2018.

[24] S. Chen, H. Xu, D. Liu, B. Hu, and H. Wang, "A vision of IoT: Applications, challenges, and opportunities with china perspective," *IEEE Internet of Things journal*, vol. 1, pp. 349-359, 2014.

[25] J. Ahmad, M. Sajjad, Z. Jan, I. Mehmood, S. Rho, and S. W. Baik, "Analysis of interaction trace maps for active authentication on smart devices," *Multimedia Tools and Applications*, vol. 76, pp. 4069-4087, 2017.

[26] I. Zualkernan, F. Aloul, S. Shapsough, A. Hesham, and Y. El-Khorzaty, "Emotion recognition using mobile phones," *Computers & Electrical Engineering*, vol. 60, pp. 1-13, 2017.

[27] G. Chetty, M. White, and F. Akther, "Smart phone based data mining for human activity recognition," *Procedia Computer Science*, vol. 46, pp. 1181-1187, 2015.

[28] X. Yang, X. Qian, and Y. Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Transactions on Image Processing*, vol. 24, pp. 1709-1721, 2015.

[29] J. S. Hare and P. H. Lewis, "Content-based image retrieval using a mobile device as a novel interface," 2005.

[30] J. Ahmad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "Saliency-weighted graphs for efficient visual content description and their applications in real-time image retrieval systems," *Journal of Real-Time Image Processing*, pp. 1-17, 2015.

[31] J. Ahmad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "Describing colors, textures and shapes for content based image retrieval-a survey," *arXiv preprint arXiv:1502.07041*, 2015.

[32] R. Mehta and K. Egiazarian, "Dominant rotated local binary patterns (DRLBP) for texture classification," *Pattern Recognition Letters*, vol. 71, pp. 16-22, 2016.

[33] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137-154, 2004.

[34] F. Jiang, Y. Fu, B. B. Gupta, F. Lou, S. Rho, F. Meng, *et al.*, "Deep Learning based Multi-channel intelligent attack detection for Data Security," *IEEE Transactions on Sustainable Computing*, 2018.

[35] J. Tang, Z. Li, and X. Zhu, "Supervised deep hashing for scalable face image retrieval," *Pattern Recognition*, vol. 75, pp. 25-32, 2018.

[36] P. Zhang, T. Zhuo, W. Huang, K. Chen, and M. Kankanhalli, "Online object tracking based on CNN with spatial-temporal saliency guided sampling," *Neurocomputing*, 2017.

[37] I. U. Haq, K. Muhammad, A. Ullah, and S. W. Baik, "DeepStar: Detecting Starring Characters in Movies," *IEEE Access*, pp. 1-1, 2019.

[38] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015, p. 6.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[40] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. Albuquerque, "Activity Recognition using Temporal Optical Flow Convolutional Features and Multi-Layer LSTM," *IEEE Transactions on Industrial Electronics*, 2018.

[41] L. Bautista-Gomez, A. Benoit, A. Cavelan, S. K. Raina, Y. Robert, and H. Sun, "Coping with recall and precision of soft error detectors," *Journal of Parallel and Distributed Computing*, vol. 98, pp. 8-24, 2016.

[42] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval* vol. 1: Cambridge university press Cambridge, 2008.

[43] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, pp. 1657-1663, 2010.

[44] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1491-1498.

[45] M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, and S. W. Baik, "Integrating salient colors with rotational invariant texture features for image representation in retrieval systems," *Multimedia Tools and Applications*, pp. 1-21, 2017.

[46] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognition Letters*, vol. 22, pp. 593-601, 2001.

[47] P. Porambage, A. Braeken, A. Gurtov, M. Ylianttila, and S. Spinsante, "Secure end-to-end communication for constrained devices in IoT-enabled Ambient Assisted Living systems," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 2015, pp. 711-714.

[48] F. S. Patel and D. Kasat, "Hashing based indexing techniques for content based image retrieval: A survey," in *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*, 2017, pp. 279-283.

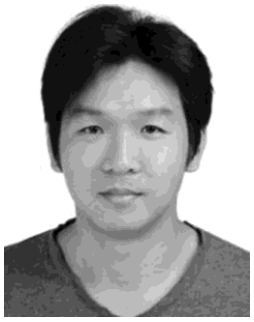
[49] Y. Li, Y. Xu, Z. Miao, H. Li, J. Wang, and Y. Zhang, "Deep feature hash codes framework for content-based image retrieval," in *Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on*, 2016, pp. 1-6.

[50] J. Ahmad, K. Muhammad, and S. W. Baik, "Medical Image Retrieval with Compact Binary Codes Generated in Frequency Domain Using Highly Reactive Convolutional Features," *Journal of medical systems*, vol. 42, p. 24, 2018.

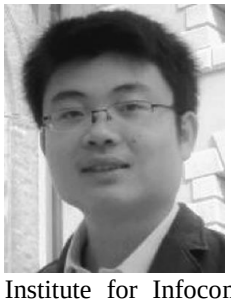
IRFAN MEHMOOD (M'16) has been involved in IT industry and academia in Pakistan and South Korea for over a decade. He is currently serving as an Assistant Professor with the Department of Software, Sejong University. His sustained contribution at various research and industry collaborative projects gives him an extra edge to meet the current challenges faced in the field of multimedia analytics. Specifically, he has made significant contribution in the areas of visual surveillance, information mining, and data encryption.

AMIN ULLAH (S'17) received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently working toward the M.S. degree leading to the Ph.D. degree in digital contents with the Intelligent Media Laboratory, Sejong University, Seoul, South Korea. His research interests include human actions and activity recognition, sequence learning, image and video analysis, and deep learning for multimedia understanding.

KHAN MUHAMMAD (S'16–M'18) received the bachelor's degree in computer science with a focus on information security from Islamia College Peshawar, Peshawar, Pakistan, in 2014, and the M.S. leading to the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2018. He is currently a Postdoctoral Researcher with the Intelligent Media Laboratory since 2018. He has authored more than 50 papers in peer reviewed international journals and conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, FUTURE GENERATION COMPUTER SYSTEMS, Neurocomputing, the IEEE ACCESS, the Journal of Medical Systems, Biomedical Signal Processing and Control, Multimedia Tools and Applications, SpringerPlus, KSII Transactions on Internet and Information Systems, MITA 2015, PlatCon 2016, FIT 2016, ICNGC 2017, and ICNGC 2018. He is an active reviewer of more than 30 reputed journals and is involved in the editing of several special issues. His research interests include information security, image steganography, video summarization, computer vision, and video surveillance.



DER-JIUNN DENG (M'10) received the Ph.D. degree from the Department of Electrical Engineering, National Taiwan University, in 2005. In August 2005, he joined the Department of Computer Science and Information Engineering, National Changhua University of Education, as an Assistant Professor and then became a Distinguished Professor in August 2016. His research interests include multimedia communication, quality-of-service, and wireless local network. Prof. Deng has received several research awards, including the Research Excellency Award of the National Changhua University of Education, the Outstanding Faculty Research Award of the National Changhua University of Education, the ICS 2014 Best Paper Award, the NCS 2017 Best Paper Award, and the Chinacom 2017 Best Paper Award. He is the Co-Editor-in-Chief for EAI Endorsed Transactions on IoT and Journal of Computers and serves as an Associate Editor for the IEEE Network Magazine.



Weizhi Meng (M'11) received the B.Eng. degree in computer science from the Nanjing University of Posts and Telecommunications, China, and obtained the Ph.D. degree in computer science from the City University of Hong Kong (CityU), Hong Kong, in 2013. He was a Research Scientist with the Infocomm Security Department, Institute for Infocomm Research, Singapore, and a Senior Research Associate with CityU. He is currently an Assistant Professor with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark. He was known as Yuxin Meng. His primary research interests are cyber security and intelligent technology in security including intrusion detection, mobile security, biometric authentication, HCI security, cloud security, trust computation, Web security, and malware analysis. He also has a strong interest in applied cryptography. He was a recipient of the Outstanding Academic Performance Award during his doctoral study and the HKIE Outstanding Paper Award for Young Engineers/Researchers in both 2014 and 2017. He was a co-recipient of the Best Student Paper Award from the 10th International Conference on Network and System Security in 2016.

FADI AL-TURJMAN received the Ph.D. degree in computer science from Queen's University, Canada, in 2011. He is a currently a Professor with Antalya Bilim University, Turkey. His record spans more than 170 publications in journals, conferences, patents, books, and book chapters, in addition to numerous keynotes and plenary talks at flagship venues. He has authored four recently published books about cognition and wireless sensor networks' deployments in smart environments with Taylor and Francis, CRC, New York (a top tier publisher in the area). He is a leading authority in the areas of smart/cognitive, wireless and mobile networks' architectures, protocols, deployments, and performance evaluation. He has received several recognitions and best papers' awards at top international conferences and led a number of international symposia and workshops in flagship ComSoc conferences. He is the Publication Chair of the 2018 IEEE International Conference on Local Computer Networks. He is serving as a Lead Guest Editor for several journals, including IET Wireless Sensor Systems, Sensors (MDPI), and Wiley.

MUHAMMAD SAJJAD received his Master degree from Department of Computer Science, College of Signals, National University of Sciences and Technology, Rawalpindi, Pakistan. He received his PhD degree in Digital Contents from Sejong University, Seoul, Republic of Korea. He is now working as an assistant professor at Department of Computer Science, Islamia College Peshawar, Pakistan. He is also head of "Digital Image Processing Laboratory (DIP Lab)" at Islamia College Peshawar, Pakistan., where students are working on research projects such social data analysis, medical image analysis, multi-modal data mining and summarization, image/video prioritization and ranking, Fog computing, Internet of Things, virtual reality, and image/video retrieval under his supervision. His primary research interests include computer vision, image understanding, pattern recognition, and robot vision and multimedia applications, with current emphasis on raspberry-pi

and deep learning-based bioinformatics, video scene understanding, activity analysis, Fog computing, Internet of Things, and real-time tracking.

VICTOR HUG C. DE ALBUQUERUE (M'17) received the graduation degree in mechatronics technology from the Federal Center of Technological Education of Ceará, Fortaleza, Brazil, in 2006, the M.Sc. degree in tele-informatics engineering from the Federal University of Ceará, Fortaleza, in 2007, and the Ph.D. degree in mechanical engineering with emphasis on materials from the Federal University of Paraíba, João Pessoa, Brazil, in 2010. He is currently an Assistant VI Professor with the Graduate Program in Applied Informatics at the University of Fortaleza, Fortaleza. He has experience in computer systems, mainly in the research fields of applied computing, intelligent systems, visualization and interaction, with specific interest in pattern recognition, artificial intelligence, image processing and analysis, Internet of Things, Internet of Health Things, as well as automation with respect to biological signal/image processing, image segmentation, biomedical circuits, and human/brain–machine interaction, including augmented and virtual reality simulation modeling for animals and humans.