Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 7586417, 12 pages
https://doi.org/10.1155/2018/7586417

Research Article

Egocentric Video Summarization Based on People Interaction Using Deep Learning

Humaira A. Ghafoor,[1] Ali Javed,[1] Aun Irtaza,[2] Hassan Dawood,[1] Hussain Dawood,[3] and Ameen Banjar[3]

[1]*Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*
[2]*Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan*
[3]*Faculty of Computing and Information Technology, University of Jeddah, Saudi Arabia*

Correspondence should be addressed to Ali Javed; ali.javed@uettaxila.edu.pk

Received 26 July 2018; Accepted 18 November 2018; Published 29 November 2018

Academic Editor: Stanislav Vítek

The availability of wearable cameras in the consumer market has motivated the users to record their daily life activities and post them on the social media. This exponential growth of egocentric videos demand to develop automated techniques to effectively summarizes the first-person video data. Egocentric videos are commonly used to record lifelogs these days due to the availability of low cost wearable cameras. However, egocentric videos are challenging to process due to the fact that placement of camera results in a video which presents great deal of variation in object appearance, illumination conditions, and movement. This paper presents an egocentric video summarization framework based on detecting important people in the video. The proposed method generates a compact summary of egocentric videos that contains information of the people whom the camera wearer interacts with. Our proposed approach focuses on identifying the interaction of camera wearer with important people. We have used AlexNet convolutional neural network to filter the key-frames (frames where camera wearer interacts closely with the people). We used five convolutional layers and two completely connected hidden layers and an output layer. Dropout regularization method is used to reduce the overfitting problem in completely connected layers. Performance of the proposed method is evaluated on *UT Ego* standard dataset. Experimental results signify the effectiveness of the proposed method in terms of summarizing the egocentric videos.

1. Introduction

The introduction of wearable cameras in 1990s by Steve Mann has revolutionized the IT industry and created a deep impact in our daily lives. The availability of low cost wearable cameras and social media has resulted in an exponential growth of the video content generated by the users on daily basis. The management of such a massive video content is a challenging task. Moreover, much of the video content recorded by the camera wearer is redundant. For example, narrative clip and GoPro cameras record a large amount of unconstrained video that contains much of the insignificant/redundant events beside the significant events. Therefore, video summarization methods [1, 2] have been proposed to address the issues associated with handling such a massive and redundant content.

Egocentric videos are more challenging to address for summarization due to the presence of jitter effects experienced because of camera wearer's movement. Accurate feature tracking, uniform sampling, and broad streaming data with very refined boundaries are the additional challenges to lifelogging video summarization. To address the aforementioned challenges associated with the egocentric videos, there exists a need to propose effective and efficient methods to generate the summary of full-length lifelogging videos. Some distinctive egocentric video recording gadgets are shown in Figure 1. The focus of these egocentric video recordings is on activities, social interaction, and user's interests. The objective of the proposed research work is to exploit these properties for summarization of egocentric videos. Egocentric video summarization has useful applications in many domains, i.e.,

Figure 1: Egocentric wearable devices [1].

law enforcement [1, 3], health care [4], surveillance [5], sports [6], and media [7, 8].

The generation and transmission of vast amount of egocentric video content in the cyberspace have motivated the researchers to propose effective video summarization techniques for wearable camera data. Existing frameworks [9–12] have used supervised as well as unsupervised methods for egocentric video summarization.

Existing methods have used supervised learning techniques for summarization based on activity detection [13–17], object detection [18], and significant events detection [19]. The goal of egocentric video synopsis is to detect significant events of the lifelogging video data and generate the summarized video. Kang et al. [20] proposed a technique to identify new objects encountered by the camera wearer. Ren et al. [21] proposed a bottom up motion-based approach to segment the foreground object in egocentric videos to improve the recognition accuracy. Hwang et al. [22] proposed a summarization technique based on identifying important objects and individuals interacted with the camera wearer. Similarly, Yang et al. [23, 24] analyzed the lifelogging video data to summarize the daily life activities of camera wearer. The summarized video contains the frames of user's interaction with important people and objects. Choi et al. [25] presented a video summarization method to identify some common human activities (i.e., talking) based on crowd perception. Lee et al. [26–28] have proposed egocentric video summarization techniques to detect the excited events of a camera wearer's entire day. These approaches [18, 26–30] have used the region cues (i.e., nearness to hands, gaze, and recurrence of event) to detect the key-events. These cues are used to evaluate the relative significance of any new region.

Egocentric video summarization methods have also used unsupervised learning to categorize sports actions [7], scene discovery [31], key-frame extraction, and summarization [29, 32]. Choudhury et al. [33] presented a pattern of influence among the people that builds on the social network. This algorithm [33] is used to find the interaction between people during the conversation. A "sociometers" wearable sensor package is used to measure face to face interaction. Yu et al. [34, 35] proposed an eigenvector analysis method to address the issue of automated face recognition. This method named as "decision modularity cut" is used to evaluate the performance in terms of social network. Fathi et al. [17] presented an approach to detect and recognize the social collaboration in a first-person video. The locality and direction information are used to compute the pattern of attention of various persons followed by assigning different roles. Finally, roles and locality information are analyzed to determine the social interactions.

In the last few years, Convolutional Neural Networks (CNNs) have been heavily explored due to its ability to learn remarkably well to understand the image content and immense scale video characterization [36–38]. Supported by the achievement of CNNs, few research works [39, 40] adopted deep learning features (e.g., CNN features) to perceive long-term activities and achieved significant implementation progress. Poleg et al. [39, 41] applied a compact 3D Convolution Neural Network (CNN) architecture for long-term activity detection of the egocentric lifelogging video data. It is a common practice to use a large and diverse dataset for CNN training in video summarization applications [42–45]. The training process has used only restricted amount of task-specific training data. Jain et al. [46, 47] used CNN features for visual detection tasks, for example, object localization, scene identification, and classification. Alom et al. [40] used cellular simultaneous recurrent networks (CSRNs) for feature extraction. CNN features computed from the supervised learning are translated-invariant.

Egocentric video recordings are inadequate with regards to a suitable structure and unconstrained in nature. Generally, there is no emphasis on the important things the user needs to record. Most important consideration in egocentric vision has concentrated on activity recognition, identification and video summarization. We proposed an effective egocentric video summarization method based on identifying the interaction of camera wearer with important people. The proposed research work aims to produce more informative summaries with minimum redundancy. The representative key-frames for the summary are selected on the basis of people interaction with the camera wearer. Our video summary focuses on the most important people that interact with camera wearer while neglecting other content. We consider interactions, such as having a discussion with the people, and fully connected with each other that are important moments. Performance of the proposed technique is evaluated on a standard egocentric video dataset "*UT Ego*" [24, 30, 48]. Experimental results show the effectiveness of the proposed method in terms of identifying important people for egocentric video summarization. Our method provides superior detection performance and generates more useful summaries with minimum redundancy as compared to existing state-of-the-arts.

The rest of the paper is organized as follows. Section 2 demonstrates the proposed framework for egocentric video summarization. Section 3 provides the results of different experiments performed on the proposed method along-with the discussion on the results. Finally, Section 4 concludes the paper.

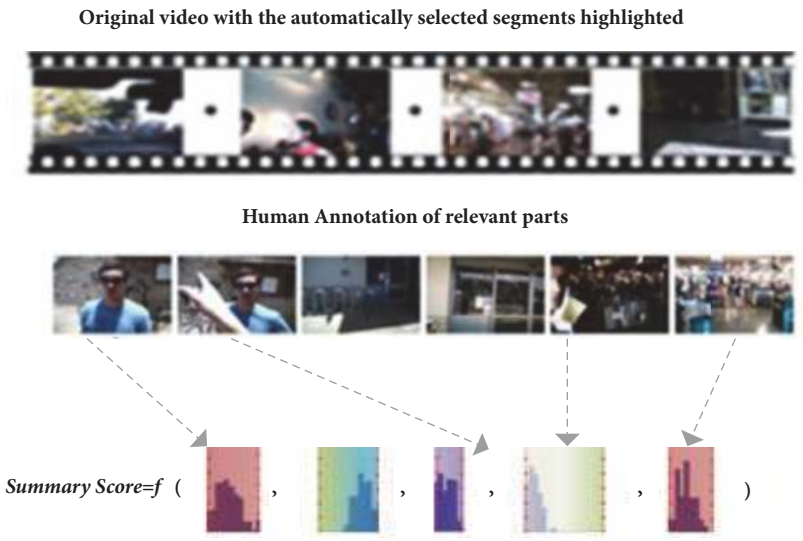


Figure 2: Summarization process of original video.

2. Materials and Methods

The proposed egocentric video summarization framework is presented in this section. Our method takes the full-length egocentric video as an input and generates a concise video that contains the most interesting segments (i.e., people interaction with camera wearer). The flow of the proposed summarization method of first-person video data is shown in Figure 2.

We trained a regression model to compute the scores of region's likelihoods. The input video is partitioned into series of n sub-shots, $V = \{s_1, \ldots, s_n\}$. We trained the AlexNet CNN model for classification. The details of our classification model are provided in Section 2.1. The architecture of the proposed framework is provided in Figure 3.

2.1. AlexNet Architecture. The proposed technique contains 8 transformation trainable layers, five convolutional layers supported by the two completely linked hidden layers and an output layer. We utilized ReLu activation function in all the trainable layers, except the last fully connected layer where we applied the Softmax function. Moreover, our system contains three pooling layers, two normalization layers, and a dropout layer. The AlexNet architecture of the proposed framework is demonstrated in Figure 4.

In the first convolutional layer, relatively large convolutional kernels of size 11 x 11 are used. For next layer, the size of convolutional kernel is reduced to 5×5, and for third, fourth and fifth layers we applied the convolutional kernels of size 3×3. In addition, first, second, and fifth convolutional layers use overlapping pooling operations with a pool size of 3×3 and stride of 2×2. Our proposed architecture has eight fully-connected layers with 4096 nodes. The last fully connected layer is supported to one thousand-way softmax function that makes dispersion over the 1000 class labels. The details of convolutional and fully connected layers are provided in Figure 5.

2.2. ReLu Non-Linearity. In recent years, the Rectified Linear Unit has recognized into a popular unit because it takes less time for training as compared to other units. Saturating nonlinearities are much slower with the non-saturating in training time with gradient descent. We used rectified linear unit function to train the network. The scope of ReLu is $[0, \text{inf}]$ which implies that it can explode the activation. The following describes the Relu activation function:

$$f(I) = \max(0, I) \tag{1}$$

where I representS the input image. If $I < 0$ the output will be zero, whereas it provides a linear function when $I \geq 0$. It is also used as a classification function. tanh is a hyperbolic tangent function that works like the sigmoid function. tanh function lies in the range of (-1,1) and computed as follows:

$$\tanh(I) = \frac{\sinh(I)}{\cosh(I)} \tag{2}$$

In this manner negative input values to the tanh will guide to negative output. The following represents the sigmoid function that lies in the range of (0,1) and computed as follows:

$$f(I) = \left(1 + e^{-I}\right)^{-1} \tag{3}$$

tanh function takes more time to train a network than ReLUs and deep convolutional neural networks. As demonstrated in Figure 6, sigmoid and tanh are computationally more complex for training purposes as compared to ReLU.

2.3. Softmax Function and Response Normalization. In the proposed architecture, we employed the softmax function as a nonlinear function at the output layer. This activation function transforms the output values into soft class possibilities. We used normalization scheme in first two layers. The activity of a neuron is computed with the aid of kernel i at position

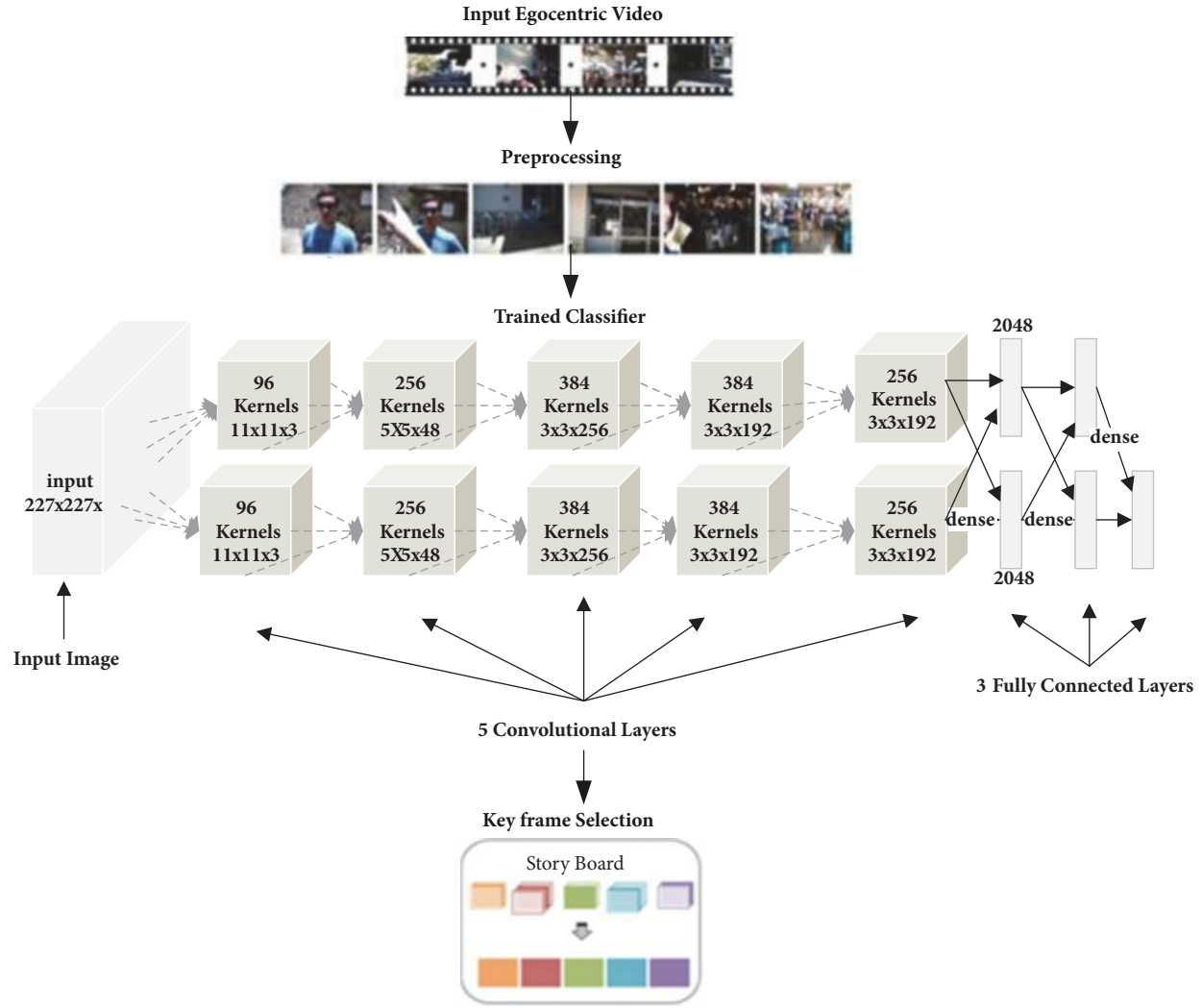


Figure 3: Architecture of the proposed egocentric video summarization framework.

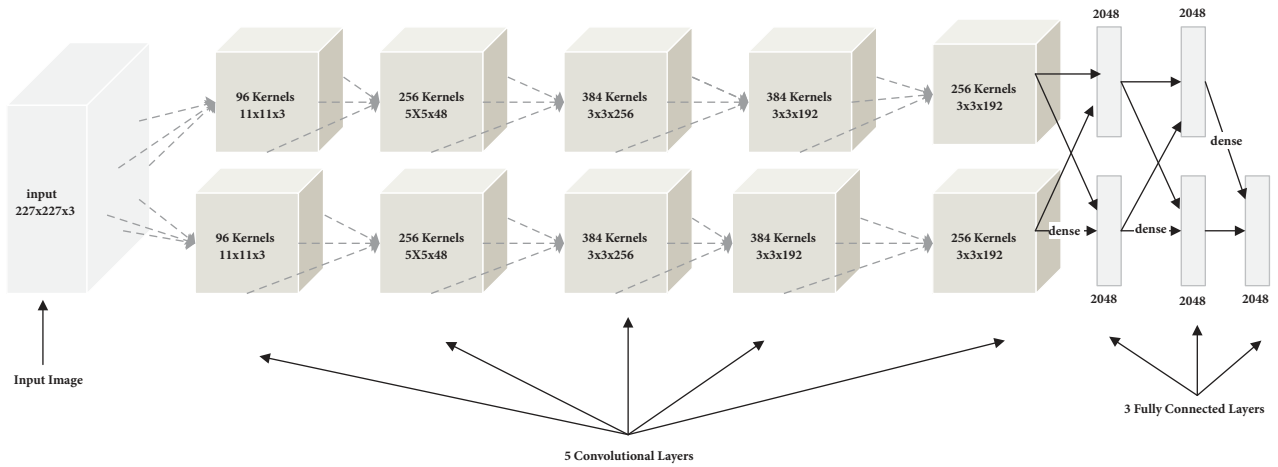


Figure 4: AlexNet architecture of the proposed method.

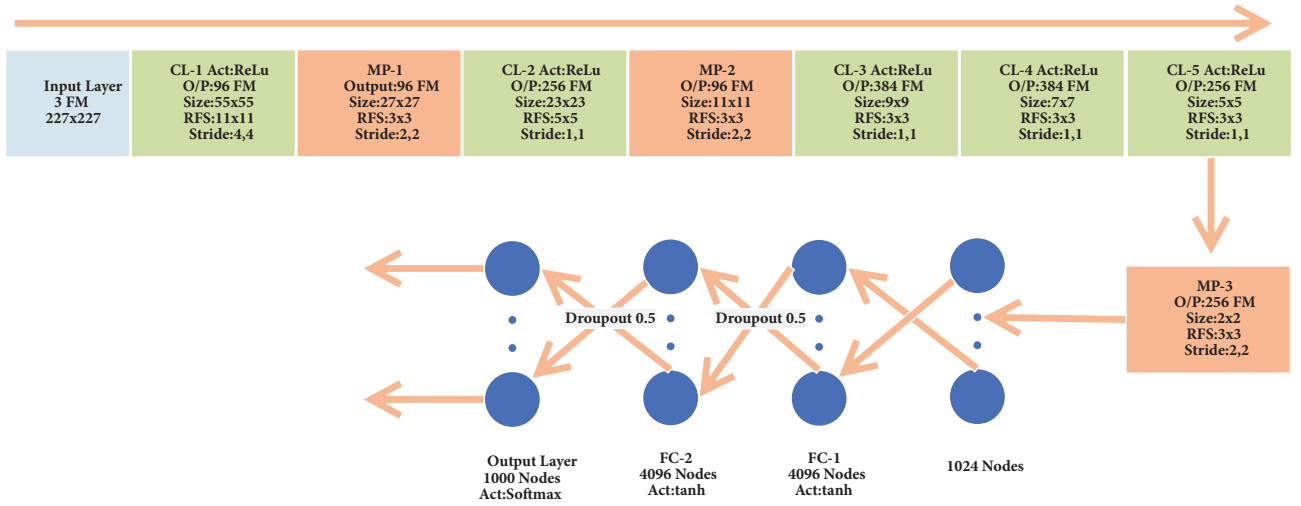


Figure 5: AlexNet architecture layers.

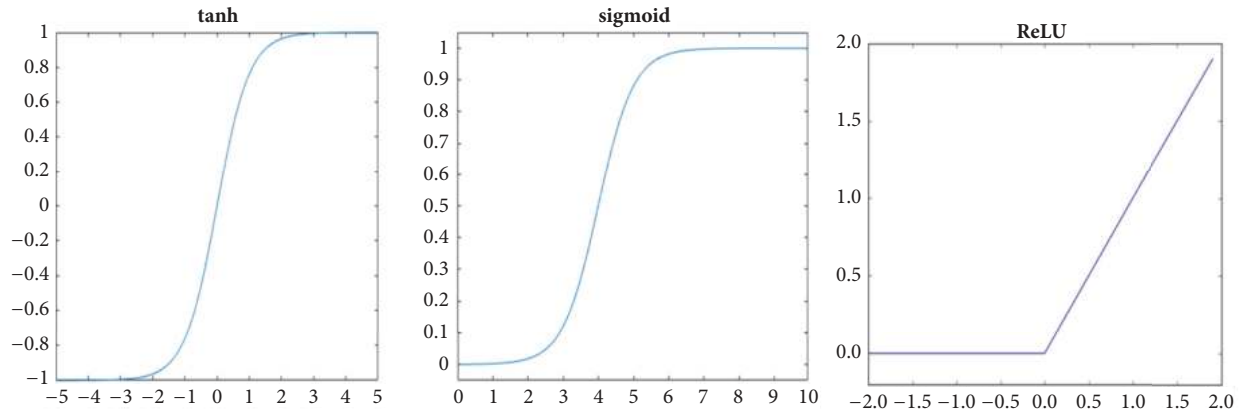


Figure 6: Comparison between tanh, Sigmoid, and ReLU training time.

(q, p) after applying the ReLU nonlinearity. The response-normalized activity is computed as follows:

$$b^{i}_{q,p} = \frac{a^{i}_{q,p}}{\left(v + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} \left(a^{i}_{q,p}\right)^{2}\right)^{\gamma}} \quad (4)$$

$a^{i}_{q,p}$ represents the activity of a neuron computed by using kernel i at position (q, p) and $b^{i}_{q,p}$ represents the response-normalized activity, where the sum runs over n "adjacent" kernels and N is the total range of kernels within the layer. The constants v, n, α, and γ are hyperparameters and their values are determined by applying a validation set. The response normalization scheme is used to reduce the test error rate of the proposed network.

2.4. Pooling Layer. We applied the overlapping pooling in the entire system. In CNN, output summary of the neighbouring groups of neurons is obtained through pooling layers in the same kernel map as these pooling units do not overlap. It requires two hyperparameters that are spatial extent w and the stride h. More specifically, this pooling layer is like a network of pooling units spaced h pixels apart. At the point of pooling unit, it summarizes a neighbourhood of size w × w. If we set h = w then we acquire the traditional local pooling as used in CNNs. By setting $h < w$, we have overlapping pooling situation where we experience lower error rate after detailed experimentation of our framework. Therefore, we used the overlapping pooling in the entire network with $h = 2$ and $w = 3$. This overlapping scheme significantly reduces the computational cost by decreasing the size of the network as well as error rate.

2.5. Dropout. When the number of iterations roughly doubles in our network, we need to converge through the dropout method. If they are "dropped out," the neurons do not participate in forward pass and back propagation. We used dropout in the initial two completely connected layers as the dropout process reduces the over fitting substantially in our proposed framework.

3. Results and Discussion

This section provides a comprehensive discussion on the results obtained through different experiments that are designed for performance analysis of the proposed framework. The details of the standard dataset used for classifier training and testing are also provided in this section. In addition, we also discussed the evaluation metrics used for measurement.

3.1. Dataset. We used a standard dataset *UT Ego* for performance evaluation of the proposed method. *UT Ego* [24, 30, 48] is specifically designed to measure the performance of egocentric video summarization approaches. *UT Ego* dataset comprises of four egocentric videos that are captured in uncontrolled environments. The dataset videos are of 3-5 hours in length having resolution of 320×480 and frame rate of 15 fps. These videos capture different daily life activities that includes eating, purchasing, attending a lecture in faculty, and driving a car. *UT Ego* dataset is divided into two classes, one where camera wearer interacts with the people and other where the camera wearer interacts with other objects.

3.2. Training and Implementation Details. The input frame is resized into 227x227 for training purposes. We used stochastic gradient descent to train our network. It has the minimum batch size of 10, momentum of 0.9, and weight decay of 0.0005 for framework to learn. The weight decay parameter value of 0.0005 reduces the error rate of our model during training. The update rule for weight ω is generated as follows:

$$m_{i+1} := 0.9.m_i - 0.0005.\epsilon.\omega_i - \epsilon.((\partial L/\partial w)|_{\omega_i})_{D_i} \quad (5)$$

$$\omega_{i+1} := \omega_i + m_{i+1} \quad (6)$$

where I, m, and ϵ represent the iteration index, momentum variable, and learning rate, respectively. $((\partial L/\partial \omega)|_{\omega_i})_{D_i}$ is the common over the ith batch D_i of the derivative of the objective with respect to ω evaluated at ω_i. The weights in each layer are introduced by zero-mean Gaussian distribution with standard deviation of 0.01. Neuron biases within the second, fourth, and fifth convolutional layers are initialized with 1. Due to this type of initialization, learning will be fast at early stages by imparting the ReLUs with fine inputs. The remaining neuron biases are initialized with 0. All layers in our network have equal learning rate which can be adjusted during the training stage. The details of training and implementation are provided in Figure 6. The learning rate was fixed at 0.01 for more reliable training and then gradually decreased to 0.0001 as the optimization stage takes more time.

3.3. Evaluation on the Validation Set. The output feature maps of our convolution layers are obtained through drop-out regularization and batch normalization. We used layer by layer dropout regularization and batch normalization. Our model will overfit if we use drop-out layer before the output layer. It has been observed that the validation set achieves better accuracy if we increase the learning features. It has to be generalized by using drop-out in each convolution layer. We used 70% images of the entire dataset for training purposes and remaining 30% for validation. Few snapshots of the training sample images, training progress, and four sample validation images along-with their predicted labels are shown in Figure 7.

3.4. Evaluation Metrics. To evaluate the performance of proposed method, three objective evaluation metrics such as precision, recall, and accuracy are used. The details of these metrics computation are provided in this subsection.

Precision represents the ratio of correctly labelled images for positive class (i.e., people interaction with camera wearer) to the total retrieved images of positive class. Precision is calculated as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

where true positive (TP) represents the frame having people interaction with the camera wearer correctly detected by the classifier. And, false positive (FP) represents the frame misclassified as positive (i.e., people interaction detected) that belongs to the negative class (i.e., frames without people interact with the camera wearer).

Recall represents the ratio of true detection of people interaction frames against the actual number of people interaction frames in the video and computed as

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (8)$$

where false negative (FN) in (8) represents the positive class images that are misclassified.

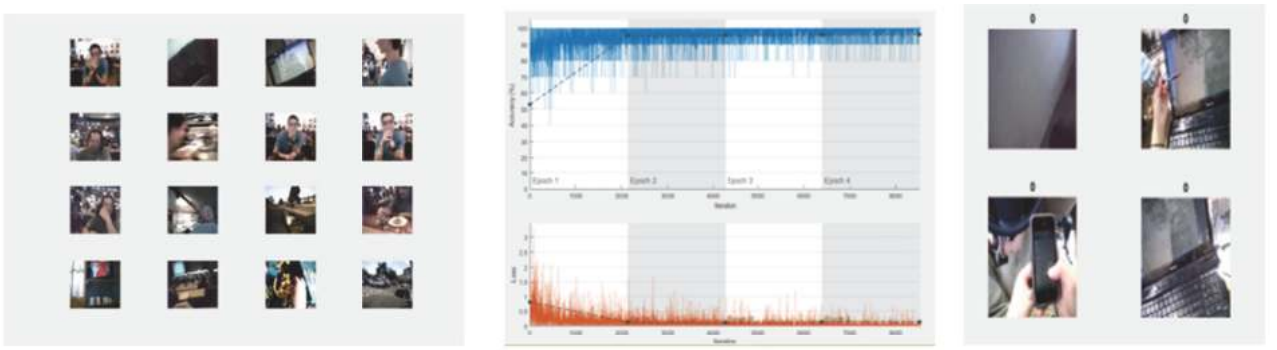
Accuracy represents the ratio of correctly labelled images of positive (i.e., images having people interact) and negative classes (i.e., images without people interaction) and computed as

$$Accuracy\ Rate = \frac{True\ Positive + True\ Negative}{Positive + Negative} \quad (9)$$

where in (9) Positive and Negative represent the total number of positive and negative samples of our dataset.

3.5. Performance Evaluation. Performance of the proposed egocentric video summarization approach is evaluated on *UT Ego* dataset. The results obtained on the *UT Ego* dataset are provided in Table 1. The proposed method obtains an average precision of 97%, recall of 95%, and accuracy of 96%. It can be clearly observed from Table 1 that the proposed approach can be used to generate informative summaries of egocentric videos. Our method performs better as it reduces the overfitting through the dropout layer and extract high level features as compared to other classifiers, i.e., SVM, KNN, etc.

3.6. Performance Evaluation of Different Classifiers for Egocentric Video Summarization. In our second experiment, we



Training Images Training Progress Predicted Labels

Figure 7: Training sample images, training progress, and sample validation images with predicted labels.

Table 1: Detection performance of proposed method.

Videos	Length (hours)	Precision	Recall	Accuracy
Video1	3:51:51	95.9	94.8	97.4
Video2	5:07:37	96.6	93.9	95.1
Video3	2:59:16	97.7	95.2	96.9
Video4	4:59:00	97.9	96.9	95.8
Average		**97**%	**95**%	**96**%

compare the overall performance of egocentric video summarization using different classifiers that are Support Vector Machine (SVM) [51], Extreme Learning Machine (ELM) [52, 53], K-Nearest Neighbor (KNN) [54], Regularized Extreme Learning Machine (RELM) [55], and Decision Trees [56]. In addition, we also compared the results obtained on these classifiers against the proposed method. The objective of this experiment is to obtain the best classification model that achieves best accuracy for egocentric video summarization based on people interaction.

We used three different feature descriptors that are Histogram of oriented gradients (HoG), local binary patterns (LBPs) and local tetra patterns (LTrPs) to train all these classifiers individually (i.e., SVM, KNN, ELM, RELM, and decision trees). More specifically, we trained each classifier (i.e., SVM) using HoG descriptor in the first phase followed by using LBP and LTrP in the second and third phase respectively. Finally, the results obtained in each phase are combined to achieve the average precision, recall, and accuracy as shown in Figure 8.

For feature extraction, we employed HoG, LBP, and LTrP on the input video frame and represent each frame in the form of feature vector for training.

For HoG descriptor representation, we decomposed input image into 64x64 sized window. A histogram of the orientated gradient is computed for each window and then normalized. The feature extraction process for HoG is shown in Figure 9.

For LBP representation, we divided the input image into small square blocks of size 3 x 3 for processing. As we know LBP is computed by comparing the centre pixel value with the neighbouring pixel values as follows:

$$LBP = \sum_{i=1}^{n} 2^{i-1} \times I\left(g^i - g^c\right) \tag{10}$$

$$I(x) = \begin{cases} 1, & x \geq 0 \\ 0, & Otherwise \end{cases} \tag{11}$$

where g^c and g^i represent the grayscale value of centre pixel and neighbouring pixels, respectively. n represents the total number of neighbours that is set to 8 in our case. Once the local binary patterns are computed for all blocks then the entire image is represented through creating the histogram as

$$H = \frac{1}{M \times N} \sum_{k=1}^{M} \sum_{l=1}^{N} f\left(LBP\left(k,l\right), q\right) \tag{12}$$

$$f(x,y) = \begin{cases} 1, & if\ x = y \\ 0, & Otherwise \end{cases} \tag{13}$$

where $M \times N$ represents the size of the image. The entire process of LBP feature extraction is demonstrated in Figure 10.

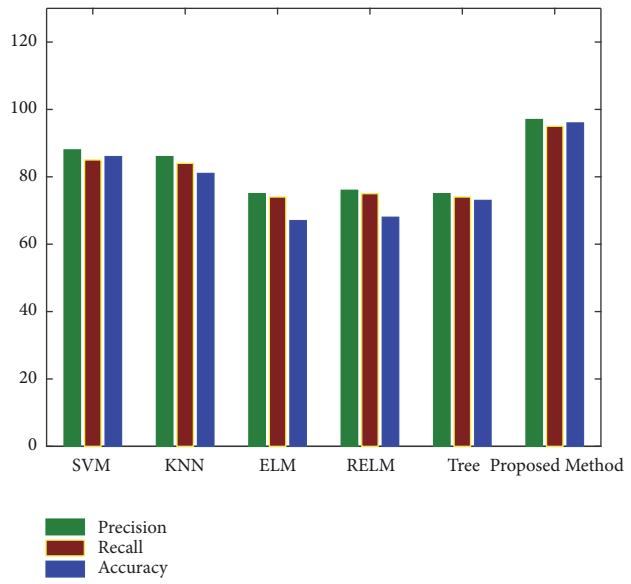


Figure 8: Accuracy, Precision and Recall rates of different classifiers.

For LTrP representation, we first resize the input image I, convert into grayscale and then calculate first-order derivatives along 0° and 90° directions as follows:

$$I^1_{0^\circ} = I\left(X_h - X_c\right)$$
$$I^1_{90^\circ} = I\left(X_v - X_c\right) \tag{14}$$

where X_h and X_v denotes the horizontal and vertical neighbourhoods of the central pixel X_c.

$$I^1_{Dir}\left(X_c\right) = \begin{cases} 1 & I^1_{0^\circ}\left(X_c\right) \geq 0 \text{ and } I^1_{90^\circ}\left(X_c\right) \geq 0 \\ 2 & I^1_{0^\circ}\left(X_c\right) < 0 \text{ and } I^1_{90^\circ}\left(X_c\right) \geq 0 \\ 3 & I^1_{0^\circ}\left(X_c\right) < 0 \text{ and } I^1_{90^\circ}\left(X_c\right) < 0 \\ 4 & I^1_{0^\circ}\left(X_c\right) \geq 0 \text{ and } I^1_{90^\circ}\left(X_c\right) < 0 \end{cases} \tag{15}$$

Depending on first order derivatives, (15) generates four directions. The values of the four directions are 1, 2, 3, and 4. Finally, a tetra bit pattern is generated by checking all the neighbouring pixels and direction of center pixel X_c. Once we obtain the LTrP we represent the entire image through histogram as shown in (12).

After representing the frames into feature vectors, we train the classifiers one by one using each of these three descriptors. We divided the dataset into two halves, the first half is used for training the classifiers and the remaining half is for testing. To be precise, we used 152165 frames each for training and validation. For SVM classification, we obtained an average precision of 88%, recall of 85%, and accuracy of 86%. For KNN, an average precision, recall, and accuracy of 86%, 84%, and 81% respectively are achieved. For ELM classification, we obtained an average precision of 75%, recall of 74%, and accuracy of 67%. For RELM classification, we obtained an average precision of 76%, recall of 75%, and accuracy of 68%. Similarly, for decision trees an average precision of 75%, recall of 74%, and accuracy of 73% are achieved. As mentioned in the previous experiment, the proposed method achieves an average precision of 97%, recall of 95%, and accuracy of 96%. From the results, it can be clearly observed that the proposed method provides superior performance as compared to SVM, KNN, decision trees, ELM, and RELM classifiers. It is concluded from the results gathered that the proposed method is very effective in terms of generating informative summaries of a full-length lifelogging video data.

3.7. Receiver Operating Characteristics Curves Analysis. In our third experiment, we designed receiver operating characteristic (ROC) curves to evaluate the performance of different classifiers along-with the proposed method. ROC curves are plotted using the false positive rate (FPR) against the true positive rate (TPR) which are computed as

$$TPR = \frac{Positive\ samples\ correctly\ classified}{Total\ positive\ samples} \tag{16}$$

$$FPR = \frac{Negative\ samples\ incorrectly\ classified}{Total\ negative\ samples} \tag{17}$$

In the proposed method, each frame is assigned a discrete class label. A (FPR, TPR) pair is obtained for each discrete classification approach that indicates a single point in ROC curve. Each point located on the curve line illustrates a pair of sensitivity and specificity values. ROC curves for SVM, KNN, decision trees, ELM, RELM, and proposed method are plotted in Figure 11. From the results we can observe that the proposed technique achieves best ROC curve among the comparative classifiers. In addition, SVM and KNN provide reasonable classification accuracy due to the fact that we have a binary classification problem. From the results we can argue that the proposed method is very effective in terms of detecting people's interaction with the camera wearer to generate more informative video summaries.

3.8. Performance Comparison of the Proposed Framework with Existing State-of-the-Art Approaches. In our last experiment, we examine the performance of the proposed method against recent existing state-of-the-art methods [23, 30, 49] for egocentric video summarization. Aghaei et al. [49] proposed a technique in the field of egocentric photo-streams captured through a low temporal resolution wearable camera. This technique [49] was deployed for multi-face detection, social signals interpretation and social interaction detection (i.e., presence or absence of people interaction). Hough-Voting for F-Formation (HVFF) and Long-Short Term Memory (LSTM) approaches were used for social interaction detection. Yang et al. [23] proposed an egocentric summarization technique for social interaction using some common interactive features like head movement, body language, and emotional expression during communication. Moreover, Hidden Markov support vector machine (HM-SVM) was used to summarize the video. Aghaei et al. [30] proposed an approach based on Long Short-Term Memory (LSTM) for detection, categorization, and social interaction of the people. A regression model was trained to detect interacting group

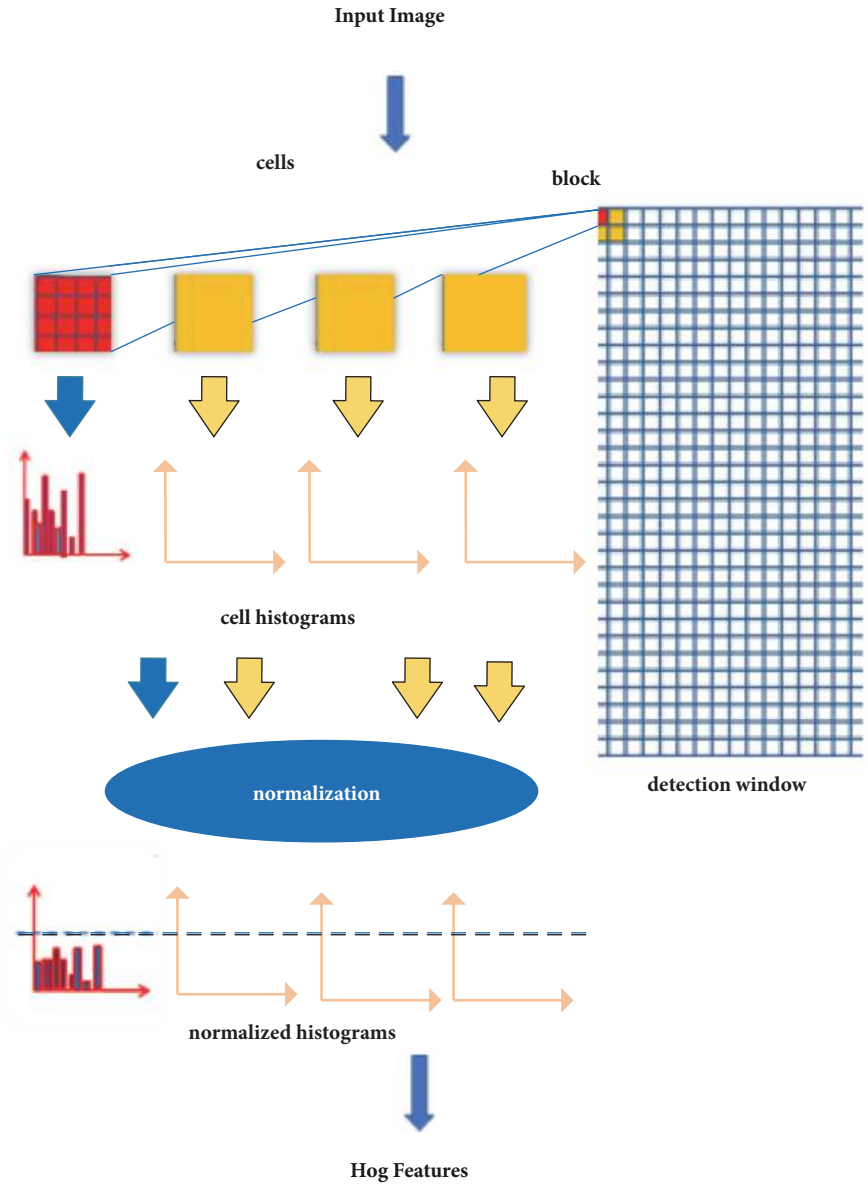


Figure 9: Histogram Oriented Gradient feature extraction process.

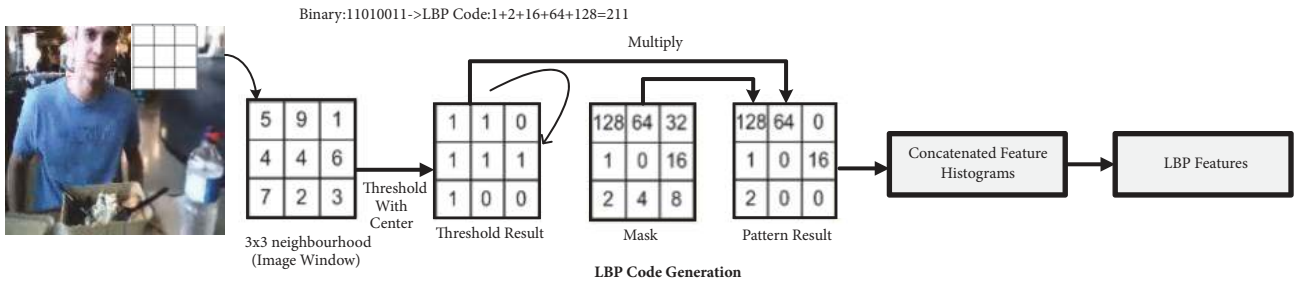


Figure 10: Local Binary Pattern feature extraction process.

Table 2: Performance comparison of the proposed and existing state-of-the-methods.

Techniques	Length (hours)	Frame Rate	Resolution	Quantity	F1-Score
Aghaei et al. [49]	Not specified	2fpm	Not specified	20.000 images	0.77
Yang et al. [23]	Not specified	30fps	Not specified	800 videos	0.91
Aghaei et al. [30]	Not specified	30fps	1920x1080	125200 images	0.87
Su et al. [50]	14	15fps	640x480	27 videos	0.82
Proposed Method	13	15fps	320x480	4 videos	0.95

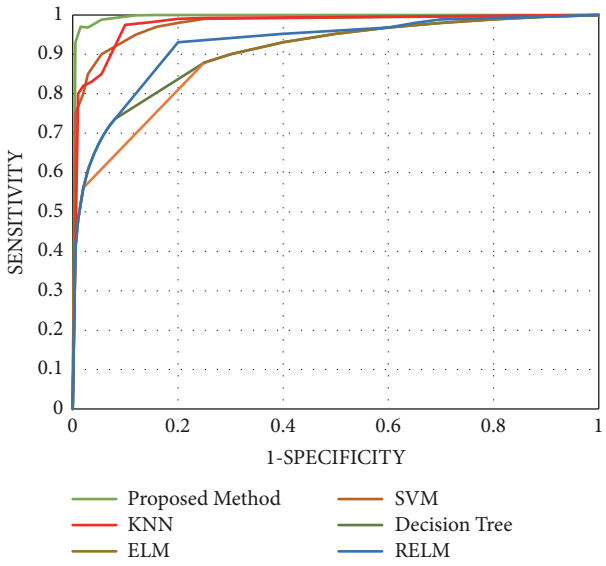


Figure 11: ROC curve analysis.

and estimate the distances between the people and camera wearer. This method [30] used low temporal resolution image sequences to detect the social interactions. Su et al. [50] proposed a video summarization approach to detect the engagement using long-term ego-motion cues (i.e., gaze). This approach [50] consists of three stages that are frame prediction, interval prediction, and classification with the trained model.

The classification performance of the proposed and comparative methods is presented in Table 2. F1-score metric is used for performance comparison as the F1-score is a reliable parameter for performance comparison in cases where some methods have better precision but lower recall and vice versa. The detailed statistics of the datasets used by each of the comparative methods are also provided in Table 2, which includes the information of video length, format, frame rate, resolution, and quantity. From Table 2, we can observe that the proposed framework shows remarkable performance and achieves encouraging results as compared with the existing methods.

4. Conclusion

In this paper, we proposed an effective method to generate the precise and informative summary of a full-length lifelogging video with minimum redundancy. Our proposed method produces the summary on the basis of people interaction with the camera wearer. The proposed scheme combines the ideas from deep convolutional neural networks and completely connected conditional random fields for key-frame extraction. The proposed method achieves an average accuracy of 96% on the challenging egocentric videos that signify the effectiveness of our method. In our experiments, we specifically used different combinations of feature descriptors on different classifiers and compared the results with our method in terms of precision, recall, and accuracy. In addition, the proposed method is also compared with existing state-of-the-art egocentric video summarization methods in terms of F1-score. Experimental results clearly indicate that the proposed technique is superior among the existing state-of-the-art techniques in terms of generating useful video summaries. Currently, we are looking to design our own egocentric video dataset with a motivation to increase the diversity of the dataset. We intend to investigate the performance of our method on a more diverse egocentric video dataset in the future.

Data Availability

The authors have used standard dataset UT Ego that is publicly available at http://vision.cs.utexas.edu/projects/egocentric_data/UT_Egocentric_Dataset.html.

Conflicts of Interest

There are no conflicts of interest. Submitting authors are responsible for coauthors declaring their interest.

Acknowledgments

This research work was performed as part of the employment of the authors under University of Engineering and Technology Taxila, Pakistan, and University of Jeddah, KSA.

References

[1] A. G. Del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of Egocentric Videos: A Comprehensive Survey," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 65–76, 2017.

[2] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multiview representative selection," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2134–2145, 2018.

[3] S. Bano, T. Suveges, J. Zhang, and S. J. Mckenna, "Multimodal Egocentric Analysis of Focused Interactions," *IEEE Access*, vol. 6, pp. 37493–37505, 2018.

[4] A. R. Doherty, K. Pauly-Takacs, N. Caprani et al., "Experiences of aiding autobiographical memory using the sensecam," *Human–Computer Interaction*, vol. 27, no. 1-2, pp. 151–174, 2012.

[5] H. Yao, A. Cavallaro, T. Bouwmans, and Z. Zhang, "Guest Editorial Introduction to the Special Issue on Group and Crowd Behavior Analysis for Intelligent Multicamera Video Surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 405–408, 2017.

[6] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am i a Baller? Basketball Performance Assessment from First-Person Videos," in *Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017*, pp. 2196–2204, Italy, October 2017.

[7] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pp. 3241–3248, USA, June 2011.

[8] C. Tan, H. Goh, V. Chandrasekhar, L. Li, and J.-H. Lim, "Understanding the nature of first-person videos: Characterization and classification using low-level features," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2014*, pp. 549–556, USA, June 2014.

[9] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 2982–2991, USA, July 2017.

[10] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 3584–3592, USA, June 2015.

[11] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 2069–2077, Canada, December 2014.

[12] B. L. Bhatnagar, S. Singh, C. Arora, and C. V. Jawahar, "Unsupervised learning of deep feature representation for clustering egocentric actions," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pp. 1447–1453, Australia, August 2017.

[13] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 2847–2854, USA, June 2012.

[14] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 407–414, Spain, November 2011.

[15] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 17–24, USA, June 2009.

[16] A. Fathi, Y. Li, and J. M. Rehg, "Learning to Recognize Daily Actions Using Gaze," in *Computer Vision – ECCV 2012*, vol. 7572 of *Lecture Notes in Computer Science*, pp. 314–327, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[17] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 1226–1233, USA, June 2012.

[18] G. Kristen, Y. L. Jae, Y.-C. Su et al., "Summarizing Long First-Person Videos. CVPR Workshop," 2016.

[19] O. Aghazadeh, J. Sullivan, and S. Carlsson, "Novelty detection from an ego-centric perspective," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pp. 3297–3304, USA, June 2011.

[20] H. Kang, M. Hebert, and T. Kanade, "Discovering object instances from scenes of Daily Living," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 762–769, Spain, November 2011.

[21] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2010*, pp. 3137–3144, USA, June 2010.

[22] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *Proceedings of the 2010 21st British Machine Vision Conference, BMVC 2010*, vol. 1, UK, September 2010.

[23] J.-A. Yang, C.-H. Lee, S.-W. Yang, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "Wearable social camera: Egocentric video summarization for social interaction," in *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo Workshop, ICMEW 2016*, 6, 1 pages, July 2016.

[24] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1059–1067, USA, July 2016.

[25] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pp. 3273–3280, USA, June 2011.

[26] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 1995–2002, Spain, November 2011.

[27] Y. Hoshen and S. Peleg, "An Egocentric Look at Video Photographer Identity," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 4284–4292, USA, July 2016.

[28] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 2235–2244, USA, June 2015.

[29] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 1346–1353, USA, June 2012.

[30] M. Aghaei, M. Dimiccoli, C. Canton Ferrer, and P. Radeva, "Towards social pattern characterization in egocentric photo-streams," *Computer Vision and Image Understanding*, 2018.

[31] N. Jojic, A. Perina, and V. Murino, "Structural epitome: A way to summarize one's visual experience," in *Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, Canada, December 2010.

[32] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. F. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proceedings of the 2008 International Conference on Image and Video Retrieval, CIVR 2008*, pp. 259–268, USA, July 2008.

[33] T Choudhury, *Sensing and modeling human networks [Ph.D. thesis]*, Massachusetts Institute of Technology, 2004.

[34] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever, "Monitoring, recognizing and discovering social networks," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 1462–1469, USA, June 2009.

[35] L. Ding and A. Yilmaz, "Learning Relations among Movie Characters: A Social Network Perspective," in *Computer Vision – ECCV 2010*, vol. 6314 of *Lecture Notes in Computer Science*, pp. 410–423, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR '14)*, pp. 1725–1732, Columbus, OH, USA, June 2014.

[37] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 568–576, Canada, December 2014.

[38] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[39] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact CNN for indexing egocentric videos," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 9, 1 pages, USA, March 2016.

[40] M. Z. Alom, M. Alam, T. M. Taha, and K. M. Iftekharuddin, "Object recognition using cellular simultaneous recurrent networks and convolutional neural network," in *Proceedings of the 2017 International Joint Conference on Neural Networks, IJCNN 2017*, pp. 2873–2880, USA, May 2017.

[41] H. Yedid and S. Peleg, "Egocentric video biometrics," *CoRR*, vol. 1, no. 3, 2014.

[42] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 1080–1088, Chile, December 2015.

[43] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 1717–1724, IEEE, Columbus, Ohio, USA, June 2014.

[44] S. Razavian, A. Hossein Azizpour, S. Josephine et al., "CNN features off-the-shelf: an astounding baseline for recognition," in *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[45] S. Pierre, D. Eigen, X. Zhang et al., *Overfeat: Integrated recognition, localization and detection using convolutional networks*, 2013, https://arxiv.org/abs/1312.6229.

[46] S. Jain, R. M. Rameshan, and A. Nigam, "Object Triggered Egocentric Video Summarization," in *Computer Analysis of Images and Patterns*, vol. 10425 of *Lecture Notes in Computer Science*, pp. 428–439, Springer International Publishing, Cham, 2017.

[47] M. M. Silva, W. L. S. Ramos, F. C. Chamone, J. P. K. Ferreira, M. F. M. Campos, and E. R. Nascimento, "Making a long story short: A multi-importance fast-forwarding egocentric videos with the emphasis on relevant objects," *Journal of Visual Communication and Image Representation*, vol. 53, pp. 55–64, 2018.

[48] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing video summarization via vision-language embedding," in *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 1052–1060, USA, July 2017.

[49] M. Aghaei, M. Dimiccoli, and P. Radeva, "Towards social interaction detection in egocentric photo-streams," in *Proceedings of the 8th International Conference on Machine Vision, ICMV 2015*, vol. 9875, Spain, November 2015.

[50] Y. Su and K. Grauman, "Detecting Engagement in Egocentric Video," in *Computer Vision – ECCV 2016*, vol. 9909 of *Lecture Notes in Computer Science*, pp. 454–471, Springer International Publishing, Cham, 2016.

[51] C. Qiong, L. Shen, W. Xie et al., "Vggface2: A dataset for recognising faces across pose and age," in *Proceedings of the Automatic Face & Gesture Recognition (FG 2018, 2018 13th IEEE International Conference on*, pp. 67–74, 2018.

[52] S. Naik and R. P. Jagannath, "GCV-Based Regularized Extreme Learning Machine for Facial Expression Recognition," in *Advances in Machine Learning and Data Science*, vol. 705 of *Advances in Intelligent Systems and Computing*, pp. 129–138, Springer Singapore, Singapore, 2018.

[53] C. Qiong, L. Shen, W. Xie et al., "Vggface2: A dataset for recognising faces across pose and age," in *Automatic Face & Gesture Recognition (FG 2018, 2018 13th IEEE International Conference on*, pp. 67–74, 2018.

[54] O. O. Khalifa, N. A. Malek, K. I. Ahmed, and F. A. Rahman, "Robust vision-based multiple moving object detection and tracking from video sequences," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, pp. 817–826, 2018.

[55] J. Tang, C. Deng, and G. B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 27, no. 4, pp. 809–821, 2016.

[56] T. Mahalingam and M. Subramoniam, "A robust single and multiple moving object detection, tracking and classification," *Applied Computing and Informatics*, 2018.