The Role of Information Fusion in Transfer Learning of Obscure Human Activities During Night

ANWAAR ULHAQ

Human actions are often tightly coupled with their context that can play an important role in their modeling and understating. However, adverse lighting conditions and clutter can easily disrupt the visual context during night, especially in outdoor environments. This situation makes it difficult for any autonomous system to detect or classify actions. Various works have proposed contextual enhancement of available imagery to improve performance. However, no study articulates the most suitable type of contextual enhancement. In this study, we try to evaluate the role of information fusion in enhancing the visual context. We are interested in knowing whether fusion can enhance the performance of the autonomous system or it is just visually appealing. Our evaluation framework is based on transfer learning using deep convolutional neural networks. Experimental results show that contextual enhancement based on 1) the fused contextual information and 2) its colorization significantly enhances the performance of automated action recognition.

Manuscript received November 20, 2019; revised February 14, 2020; released for publication June 30, 2020.

The author is with the School of Computing and Mathematics, Charles Sturt University, Port Macquarie, NSW 2444, Australia (E-mail: aulhaq@csu.edu.au).

1557-6418/20/$17.00

I. INTRODUCTION

Human action recognition is a challenging computer vision problem. Different challenging scenarios are considered in the literature like action by large groups [1], group actions [2], recognizing actions in crowd [3], actions inside movies [4], single- and two-person action recognition [5], action recognition from the side of a video [6], actions across different viewpoints [7]–[9], and occluded actions [10]. However, these approaches assume that the action dataset is captured at daytime under clear context and reasonable lighting conditions. Their performance will decline if available data are adversely affected by diverse lighting conditions and cluttered context. Fig. 1 illustrates two such scenarios of adverse lighting conditions at night. Targets (actors) are visible with dim and hazy context in infrared (IR) imagery. In contrast, background context is clear with hidden or vague targets in visible (VIS) imagery. In this paper, we want to explore if context is enhanced, how it contributes to the automated recognition through machine vision.

Visual context is valuable a priori knowledge in terms of modeling action instances. Therefore, contextual action recognition is addressed by various researchers. The context of scenes is utilized for recognizing events by Li and Fei-Fei [11]; however, it uses only the static images. Contextual action recognition is presented by Marszalek et al. [12], which is based on the bag-of-features framework. It considered the annotated actions in movies and with script mining for visual learning. A similar technique [13] extracts the overall object-based context by detectors and their descriptors with supervised learning. Modeling of scene and object context is designed by Jiang et al. [14] for the Hollywood2 action dataset. These approaches aim at action recognition in high-resolution videos. Hierarchical attention and context modeling for group activity recognition is considered in recent works [15], [16]. However, achieving the same objectives in night-time imagery is cumbersome due to clutter and low-lighting conditions.

Human action activity recognition in a single spectrum is discussed in [17] and [18], which perform recognition in IR spectrum. However, these approaches ignore action contexts that are poorly captured by IR sensors. These approaches, therefore, cannot be classified as contextual action recognition approaches. In this paper, we build upon the idea of [19]–[21] and further evaluate the role of contextual information fusion in recognizing human actions.

Moreover, night-time imagery lacks color information that provides great help to human visual perception. Due to unnatural appearance and IR imagery limitations, multi-sensor systems and color information are integrated for better contextual awareness [22]. Another approach to optimize these systems is to introduce pseudo-color information [23], [24]. A recent study about the perceptual evaluation [25] of such

Fig. 1. Two different scenarios of visual context of actions captured by two different sensors: low-light VIS and IR sensors. Each sensor has its limitation that affects recognition performance. It means contextual improvement can play a positive role in improvement of detection capability of autonomous systems.

color-transformed multispectral systems concludes that pseudo-colorization better illustrates the gist of a night scene by improving the fixation behavior of human eye compared to large-scale imagery.

We address the following research question: Can accuracy of automated action recognition be increased by context enhancement through information fusion and transferring knowledge from daytime image data to night-time data?

This paper claims the following contributions: 1) It evaluates the role of information fusion in transfer learning of activity recognition at night-time. To the best of our knowledge, it is the first work that evaluates such a problem. 2) It explores how transfer learning can be better utilized (frozen or fine-tuned) for transferring knowledge from different domains.

The paper is organized as follows: Section 2 presents the related work, Section 3 illustrates how the context enhancement of multisensor videos is possible, and Section 4 discusses the transfer learning framework and the action filter. Experimental results are discussed in Section 5. The conclusion and references are provided at the end.

II. PRIOR WORK

Human action recognition is now a well-researched area. There are various methodologies that can be categorized on the basis of the scenario used. The performances of these approaches vary in different circumstances and challenges. One of such challenges is the action context. An action-scene context is acquired through movie-script mining by Liu et al. [1] for realistic action recognition in movies. Spatiotemporal action context was utilized by Han et al. [13] based on space–time features. Similarly, [26] employs convolutional neural networks (CNNs) for contextual action recognition. However, these approaches use high-resolution action datasets for which the extraction of spatiotemporal interest points is straightforward.

Recently, deep CNNs [27] have achieved significant success in object detection and classification. In particular, CNNs trained on the large datasets such as ImageNet have been shown to learn general-purpose image descriptors for a number of vision tasks. A recent trend has been observed about the use of deep feature learning. Various pretrained convolutional network (ConvNet) models are publicly available. In the same spirit, 3D ConvNets [6], [28] were proposed for different types of video analysis tasks, especially action recognition. Instead of using fully connected layers, activations from convolutional layers of the network have achieved superior results.

However, recognition of human actions in low-quality night-time videos is not well-explored area of research and very few approaches can be cited in this category. The utility of thermal imagery is analyzed by Li and Gong [29] for human action recognition. This approach is built upon the histogram of oriented gradients and nearest-neighbor classification.

A similar work [30] uses gait energy images. However, it is limited to walking activity, which is easier to recognize. IR image super-resolution is proposed for enhancement by Du et al. [31]. Deep VIS and thermal image fusion [32] was used for enhanced pedestrian visibility. However, such work cannot be categorized as action recognition work. Fourier transform is a great tool to analyze response of patterns of interest in the frequency domain. Such matching is efficient and faster than matching based on spatial templates. In addition, it combines target classification and detection (localization) simultaneously. Inspired by this this idea, a contextual action recognition approach based on 3D fast Fourier transform and contextual cues was proposed in [19] and [20].

In this paper, we present robust action recognition, which can deal with low-quality night-time video sequences. In case of night-time videos, we consider the registered videos collected from low-light VIS and IR spectra. We enhance the context through video fusion [33]. Our action recognition approach is based on space–time interest point detection and frequency-domain correlation analysis and can detect and classify human actions in a robust manner.

III. CONTEXT ENHANCEMENT OF NIGHT-TIME VIDEOS

In this section, we discuss the motivation behind contextual enhancement of night-time video sequences, video fusion, and colorization for context enhancement.

A. Motivation

The aim of context enhancement is a preprocessing step to give day-like appearance to night-time videos.

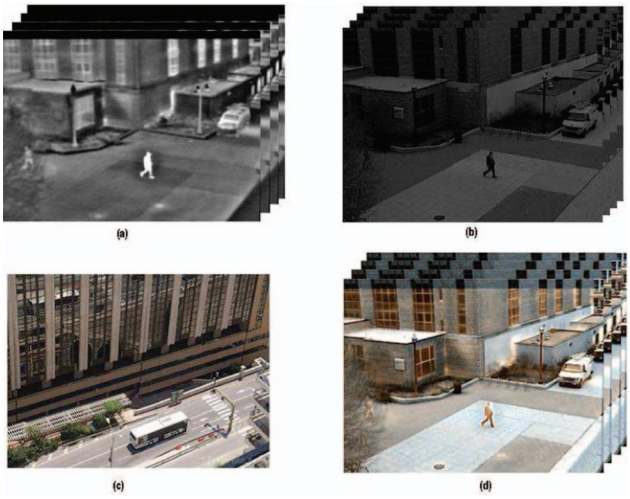

Fig. 2. Color-transfer-based video fusion method: (a) an IR video sequence; (b) low-light VIS domain video sequence; (c) a source color image for the purpose of color transfer; and (d) a color-fused video generated from (a), (b), and (c). Image adapted from [22].

It involves video fusion applied on registered video streams collected from IR and VIS spectra. Context enhancement helps to reveal a camouflaged target and to assist target localization [34]. Here, we present and discuss context enhancement briefly.

B. Context Enhancement Through Video Fusion and Colorization

The objective of employing video fusion is to generate a single enhanced video from complementary videos that is more suitable for the purpose of human visual perception, action, and context recognition. If we denote A as IR video sequence and B as a VIS video sequence, we intend to generate another video sequence C by fusing visual information from A and B. Fig. 2 gives illustrations of video fusion results.

There is extensive literature on contextual enhancement of nigh-time imagery. However, we selected automatic color-transfer-based video fusion (FACE) [22] as it enhances video context by color transfer from a source image because it enhances context through fusion and colorization simultaneously. An example scenario adapted from this work can be seen in Fig. 2.

IV. MULTIDOMAIN ACTION RECOGNITION VIA TRANSFER LEARNING

Transfer learning is a machine learning methodology where a model developed for a task in one domain is reused as the starting baseline for learning a more specific model on another task in the other domain. Let us define domain and task for better understanding of transfer learning.

Let D denote the domain; we can define it as a two-element tuple (a finite ordered list) consisting of feature space, X, and marginal probability, $P(X)$, where X is a sample data point. Therefore, we can write the domain mathematically as $D = (X, P(X))$.

A task, T, then can be defined as a two-element tuple of the label space, Y, and objective function, O, denoted as $P(Y|X)$ from a probabilistic viewpoint. Thus, a task can be defined as $T = (Y, P(Y|X)) = (Y, O)$. We can define transfer learning as follows.

Given a source domain D_s with source task T_s, and similarly, a target domain D_t with target task T_t, transfer learning has the objective to learn the target conditional probability distribution $P(Y_T|X_T)$ in D_T with the knowledge transferred from D_s and T_s, where $D_s \neq D_t$ and $T_s \neq T_t$. Usually in such scenarios, the number of labeled target examples is exponentially smaller than the number of labeled source examples. We have a similar scenario as the majority of action datasets and trained models have daytime-captured data, while target night-vision data are scarce. Our problem in this study, however, is not the design of effective transfer learning but to evaluate the suitability of data for transfer learning.

We will use transfer learning to extract knowledge from the already trained 3D CNN [2] for action recognition in one domain (daytime) and would use it to learn actions in the other domain (night time). This pretrained network on the UCF101 action dataset has eight convolutions, five maximum pooling, and two fully connected layers, followed by a softmax output layer.

All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. After training all the network layers, we extract fc6 layer features from the trained network and call them C3D features. To extract these features, we follow the guidelines of [2]. Each video is split into clips, and each of the 16 frames is passed to the C3D network to extract fc6 activations. These fc6 activations are averaged to form a 4096-dimensional video descriptor followed by an L2 normalization. These representations are known as C3D video features.

In transfer learning, we will train a base network and then copy its first n layers to the first n layers of a target network. In our case, $n = 4$. The remaining layers of the target network will then randomly be initialized and trained toward the target task. Rather than freezing the transferred layers, we train all the layers. It is inspired by the recent work by Yosinski et al. [35].

A. Action Classification

To apply transfer learning for the action classification task, we developed four separate networks based on the 3D CNN model, C3D network [2]. The first network was trained on daytime image data UCF101 [36]. Model A is 3D CNN trained on IR video sequences (InfAR data, and other available night IR data), Model B is 3D CNN trained on the night-time VIS spectrum only, and

Model C is 3D CNN trained on fused video sequences and color-enhanced video sequences.

B. Action Detection

Action detection is more challenging compared to simple classification as it not only classifies the action but also provides its location. To achieve action detection, we use 3D feature-based zero-aliasing maximum-margin correlation filter as described below.

1) Action-02MCF: 3D Feature-Based Zero-Aliasing Maximum-Margin Correlation Filter:

The motivation of using correlation filter compared to end-to-end classification is simultaneous localization and detection of action instances. We train 3D correlation filters on fine-tuned features described in the previous section. These filters can be synthesized by calculating Fourier transfer of fine-tuned features. Correlation filters were initially developed in the seminal work of [37], which is a way of learning a template/filter in the frequency domain that, when correlated with a set of training signals, gives a desired response (correlation peak). A general correlation filter h can be expressed as

$$h = arg_h min \sum_{i=1}^{N} ||h \otimes x_i - g_i||^2, \tag{1}$$

where $\otimes$ denotes the cross-correlation of the vector versions of the input signal x_i and the template h, and g_i is the vector version of the desired correlation output. If N denotes the training feature vectors, x_i denotes the ith feature vector.

Correlation filters are generally 2D as these filters work well on images. In our previous work [38], we have extended correlation filters in 3D for action recognition. Therefore, only a brief description of their optimization criteria is presented here as the complete design of Action-02MCF correlation filters is described in [38].

The correlation filter design problem is often considered as an optimization problem. If N denotes the training feature vectors of length M, we can write the multiobjective function of the proposed correlation filter as follows:

$$\begin{aligned} h = \min_{\hat{h}} \Bigg(& \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{M} || \hat{f}_i^k \otimes \hat{h}_i^k \\ & - \hat{g}_i||_2^2, \lambda \sum_{k=1}^{M} ||h^k||_2^2 + C \sum_{i=1}^{N} \xi_i \Bigg), \\ s.t. \quad & y_i \left(\sum_{k=1}^{M} \cdot \hat{h}^{k'} \cdot \hat{f}_i^k \right) \geqslant u_i \xi_i \end{aligned}$$

Here, $\hat{f}$ is a feature vector, $\hat{h}$ is a frequency-domain filter, ξ_i is a penalty term to penalize the training samples on the wrong side of the margin, λ is a regularization parameter, whereas $C > 0$ denotes a trade-off parameter, y_i is a class label (1: positive class; -1: negative class), and u_i is the minimum peak value set to 1 for N training samples and an M number of features.

The regularization parameter λ serves as a degree of importance that is given to misclassifications. So, intuitively, the larger the λ grows, the fewer the wrongly classified examples are allowed (or the higher the price they pay in the loss function). Then, when λ tends to infinity, the solution tends to the hard margin (allowing no misclassification). When λ tends to 0 (without being 0), more misclassifications are allowed.

C parameter controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights and it tells the optimization how much misclassification to be avoided for each training example. For large values of C, the optimization will choose a smaller margin hyperplane if that hyperplane does a better work of achieving all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger margin separating hyperplane, even if that hyperplane misclassifies more points.

2) Notation: Vectors are represented by the lowercase letters f, matrices are represented by uppercase letters F, ˆ represents variables in the frequency domain, and t represents its transpose.

We extracted C3D features and named them Model-0-fc. We call these models as the pretrained model and extracted features as pretrained features. Model-A-fc was extracted from 3D CNN trained on IR video sequences (InfAR data, and other available night IR data), Model-B-fc was extracted from 3D CNN trained on the night-time VIS spectrum only, and, finally, Model-C-fc was extracted from 3D CNN trained on fused video sequences and color-enhanced video sequences.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section describes our experimental data, setup, results, and performance comparison with discussion.

A. Action Dataset and Experimental Setup

In the absence of any benchmark night-vision (NV) action dataset, we have recorded the NV action dataset using two different cameras. One of them is an IR camera, Raytheon Thermal IR-2000B, and the other is a low-light VIS camera, Panasonic WV-CP470. The thermal and visual videos are registered before the fusion process. In addition to these videos, this dataset includes 20 video sequences collected from the TNO image fusion dataset [25], Eden Project dataset [39], and Ohio State University thermal dataset. This dataset comprises eight action categories, including walking, wave1, wave2, stand-up, sit-down, clapping, pick-up, and running performed by different actors. It also includes videos from the IR action dataset [40], which contains 12 common human actions with IR video sequences. All action

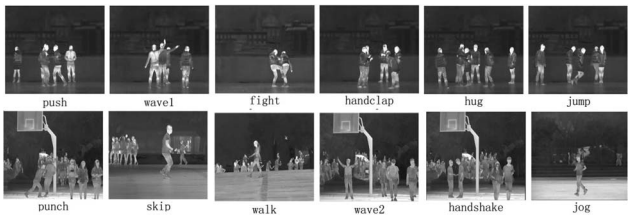

Fig. 3. Sample IR video instances for 12 action classes. Action categories include fight, handclasp, handshake, hug, jog, jump, punch, push, skip, walk, wave1, and wave2.

instances are displayed in Fig. 3. Sample actions include one-hand wave (wave1), multiple-hand wave (wave2), handclasp, jog, jump, walk, skip, hug, push, handshake, punch, and fighting action, each containing 50 video sequences, with 25 fp/s and resolution of 293 × 256. These actions are performed by 40 different actors. Each video clip lasts about 4 s on average. Some of these videos illustrate interactions between multiple actors.

B. Experiment No. 1: The Role of Information Fusion in Terms of Error Rate

In this experiment, we intended to validate the significance of multisensor fusion data. We created training and validation sets with a ratio of 70:30. The base network was trained on IR data only as actions are more visible in the IR than in the VIS spectrum. The second network is prepared after transferring the first three layers of the C3D network and remaining layers of the fused dataset. Error rate is calculated for both training and validation sets. The experiment is shown in Fig. 5. It shows that in the case of fused data, the error rate is much lower than that of single-domain data for both training and validation sets.

C. Experiment No. 2: The Role of Information Fusion in Terms of Recognition Accuracy

This experiment checks the classification accuracy. For validation, the leave-one-out cross-validation strategy is used. The results are shown in Table 1 in terms of recognition accuracy and the layers used during transfer learning. We experimented with different versions of our recognition framework to know the impact of information fusion on recognition performance. First, a baseline is developed as discussed in the transfer learning section. It is based on the daytime video action dataset, and we used the knowledge extracted from this network to fine-tune other networks. This network was the C3D network pretrained on the UCF101 dataset. Second, we fine-tuned other networks as described earlier in the transfer learning section. In addition, to know the effect of different layers in transfer learning, we fine-tuned different versions of each network to quantify the learning transferred from the base network.

Table 1
Average Recognition Accuracy of Three Different Models Versus Number of Layers Transferred from the Baseline Network

Model used	No. of layers transferred	Recognition accuracy
Model A	3, 4, 5	0.96, 0.91, 0.87
Model B	3, 4, 5	0.72, 0.71, 0.69
Model C	3, 4, 5	0.98, 0.93, 0.87

It shows that best recognition for each model is achieved if only three convolution layers are transferred as after this dataset specificity started increasing.

In this experiment, Model A is 3D CNN trained on IR video sequences (InfAR data, and other available night IR data), Model B is 3D CNN trained on the nighttime VIS spectrum only, and Model C is 3D CNN trained on fused video sequences and color-enhanced video sequences.

We calculated the average recognition accuracy for each case against the NV dataset and the results are displayed in Table 1. We found that network that uses both color and context information fusion alongside motion cues outperforms others. It demonstrated that fused information is significantly important in the action recognition process. Contextual information also plays an important role, especially in actions that involve full-body motion. Therefore, an information fusion of motion, color, and contextual cues can enhance action recognition performance.

D. Experiment No. 3: Filter Performance for Action Detection and Localization

For this experiment, the training of Action-02MCF filters is performed for each action category. During the testing phase, a test action video is correlated with the synthesized filter to find the correlation peak.

To measure the detection and localization performance of the proposed filter, we utilize the probability of detection versus false alarms per second (FA/s). A performance metric denoted as P is utilized, which is equal to the integration of a receiver operating characteristic (ROC) curve from 0 to 5 FA/s. An ideal ROC curve must have $P = 5$. To evaluate this performance,

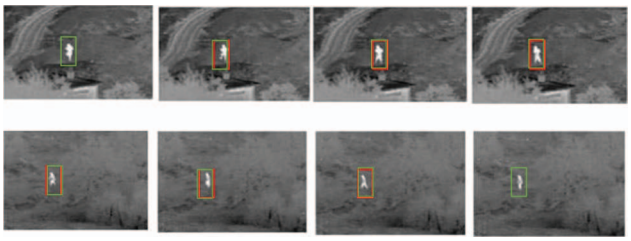

Fig. 4. Action instance detection in three NV action instances, where the red bounding box is the actual ground truth, while the green bounding box shows detection by 3D SDCF.

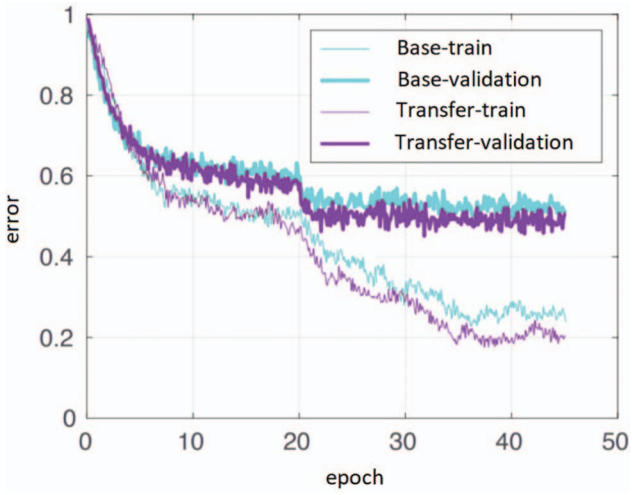

Fig. 5. The plot for the effect of transfer learning on the C3D network. The base C3D network trained on the NV dataset shows a high error rate in both training and validation sets compared to the C3D transferred network that is fine-tuned on the knowledge transferred by the fused dataset.

we applied the proposed filter to each test video and varied the threshold of the detection to generate ROC curves. The detection is labeled a true positive detection if the ground truth and the center of the bounding box lie within three frames of each other and the Euclidean distance is $\leq$8 pixels in the spatial domain to keep a >50% bounding box overlap for each action. We then plot the values of the performance metric P against all actions and perform a comparison with a similar approach,

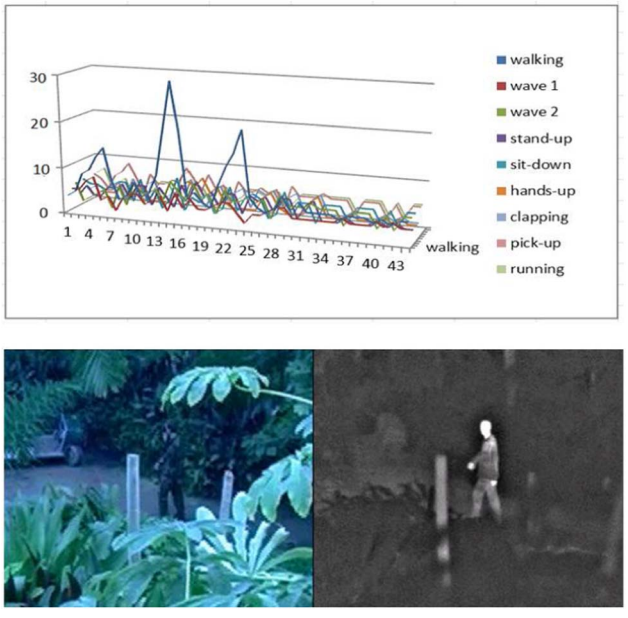

Fig. 6. Top: the plot of PSR by correlating the trained walking-Action-02MCF with a night-time video sequence. As visible from the plot, PSRs produced by walking-Action-02MCF are comparatively much higher than the responses by action filters for other actions. Bottom: a representative frame from the respective videos (both domains).

Action-DCCF filter [19]. The corresponding filter is selected due to similarity and code availability. Fig. 4 displays the sample detections with the original and estimated bounding boxes.

E. Experiment No. 4: Quantitative Evaluation of Filter Robustness

We use another quantitative metric, named peak-to-sidelobe ratio (PSR) described in [41], which calculates the ratio of peak response to local surrounding response. Fig. 6 plots PSRs for walking action present in the test video sequence (sample frame displayed) using Action-03MCF for the walking action trained on the NV dataset.

VI. CONCLUSION

In this paper, we explored and discussed the role of information fusion for automated action recognition. We use deep ConvNets for action recognition and used transfer learning to learn and transfer knowledge from a pretrained action network. In addition, we included an action-detection framework based on robust feature-based space–time action recognition called Action-02MCF. Experiments were conducted to know the effects of transfer learning, number of layers in transfer learning, and information fusion on improving the performance of action recognition at night time. We discovered that information fusion enhances action recognition performance as it improves the contextual information of night-vision data.

REFERENCES

[1] J. Liu, J. Luo, and M. Shah
"Recognizing realistic actions from videos 'in the wild'," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1996–2003.

[2] S. Gong and T. Xiang
"Recognition of group activities using dynamic probabilistic networks," in
Proc. IEEE Int. Conf. Comput. Vis., 2003, pp. 742–749.

[3] P. Siva and T. Xiang
"Action detection in crowd," in
Proc. Brit. Mach. Vis. Conf., 2010, pp. 1–11.

[4] I. Laptev, M. Marsza?ek, C. Schmid, and B. Rozenfeld
"Learning realistic human actions from movies," in
Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2008, pp. 1–8.

[5] S. Islam, T. Qasim, M. Yasir, N. Bhatti, H. Mahmood, and M. Zia
"Single- and two-person action recognition based on silhouette shape and optical point descriptors,"
Signal Image Video Process., vol. 12, no. 5, pp. 853–860, 2018.

[6] L. Pei, M. Ye, X. Zhao, T. Xiang, and T. Li
"Learning spatio-temporal features for action recognition from the side of the video,"
Signal Image Video Process., vol. 10, no. 1, pp. 199–206, 2016.

[7] A. Ulhaq, X. S. Yin, J. He, and Y. Zhang
"On space–time filtering framework for matching human actions across different viewpoints,"

IEEE Trans. Image Process., vol. 27, no. 3, pp. 1230–1242, 2018.

[8] A. Ulhaq, I. Gondal, and M. Murshed
"On dynamic scene geometry for view-invariant action matching," in
Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 3305–3312.

[9] A.-U. Haq, I. Gondal, and M. Murshed
"On temporal order invariance for view-invariant action recognition,"
IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 2, pp. 203–211, 2012.

[10] D. Weinland, M. Özuysal, and P. Fua
"Making action recognition robust to occlusions and viewpoint changes," in
Computer Vision–ECCV. Berlin, Germany: Springer, 2010, pp. 635–648.

[11] L.-J. Li and L. Fei-Fei
"What, where and who? Classifying events by scene and object recognition," in
Proc. IEEE 11th Int. Conf. Comput. Vis., 2007, pp. 1–8.

[12] M. Marszalek, I. Laptev, and C. Schmid
"Actions in context," in
Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 2929–2936.

[13] D. Han, L. Bo, and C. Sminchisescu
"Selection and context for action recognition," in
Proc. Int. Conf. Comput. Vis., 2009, vol. 9, pp. 1933–1940.

[14] Y.-G. Jiang, Z. Li, and S.-F. Chang
"Modeling scene and object contexts for human action retrieval with few examples,"
IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 5, pp. 674–681, 2011.

[15] L. Kong, J. Qin, D. Huang, Y. Wang, and L. Van Gool
"Hierarchical attention and context modeling for group activity recognition," in
Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 1328–1332.

[16] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes
"Hierarchical attention network for action segmentation,"
Pattern Recognit. Lett., vol. 131, pp. 442–448, 2020.

[17] J. Han and B. Bhanu
"Human activity recognition in thermal infrared imagery," in
Proc. IEEE Comput. Vis. Pattern Recognit. Workshop, 2005, pp. 17–17.

[18] J. F. Li and W. G. Gong
"Application of thermal infrared imagery in human action recognition,"
Adv. Mater. Res., vol. 121, pp. 368–372, 2010.

[19] H. Anwaar, G. Iqbal, and M. Murshed
"Contextual action recognition in multi-sensor nighttime video sequences," in
Proc. Digit. Image Comput. Techn. Appl., 2011, pp. 256–261.

[20] A. Ulhaq
"Action recognition in the dark via deep representation learning," in
Proc. IEEE Int. Conf. Image Process., Appl. Syst., 2018, pp. 131–136.

[21] G. I. Haq Anwaar and M. Manzur
"Action recognition using spatio-temporal distance classifier correlation filter," in
Proc. Int. Conf. Digit. Image Comput. Techn. Appl., 2011, pp. 474–479.

[22] A. Ulhaq, X. Yin, J. He, and Y. Zhang
"FACE: fully automated context enhancement for night-time video sequences,"
J. Vis. Commun. Image Representation, vol. 40, pp. 682–693, 2016.

[23] A. Ulhaq, I. Gondal, and M. Murshed
"Scarf: semi-automatic colorization and reliable image fusion," in
Proc. Int. Conf. Digit. Image Comput. Techn. Appl., 2010, pp. 435–440.

[24] A. Ulhaq, A. Mirza, and S. Qamar
"An optimized image fusion algorithm for night-time surveillance and navigation," in
Proc. IEEE Symp. Emerg. Technol., 2005, pp. 138–143.

[25] A. Toet, M. J. de Jong, M. A. Hogervorst, and I. T. Hooge
"Perceptual evaluation of colorized nighttime imagery," in
Proc. IS&T/SPIE Electron. Imaging, 2014, pp. 901412–901412.

[26] G. Gkioxari, R. Girshick, and J. Malik
"Contextual action recognition with R*CNN," in
Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1080–1088.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton
"Imagenet classification with deep convolutional neural networks," in
Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[28] S. Ji, W. Xu, M. Yang, and K. Yu
"3D convolutional neural networks for human action recognition,"
IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, 2013.

[29] J. F. Li and W. G. Gong
"Application of thermal infrared imagery in human action recognition,"
Adv. Mater. Res., vol. 121, pp. 368–372, 2010.

[30] J. Han and B. Bhanu
"Human activity recognition in thermal infrared imagery," in
Proc. IEEE Comput. Vis. Pattern Recognit. Workshop, 2005, pp. 17–17.

[31] J. Du, H. Zhou, K. Qian, W. Tan, Z. Zhang, L. Gu, and Y. Yu
"RGB-IR cross input and sub-pixel upsampling network for infrared image superresolution,"
Sensors, vol. 20, no. 1, p. 281, 2020.

[32] I. Shopovska, L. Jovanov, and W. Philips
"Deep visible and thermal image fusion for enhanced pedestrian visibility,"
Sensors, vol. 19, no. 17, p. 3727, 2019.

[33] H. Anwaar, G. Iqbal, and M. Murshed
"Automated multi-sensor color video fusion for nighttime video surveillance," in
Proc. IEEE Symp. Comput. Commun., 2010, pp. 529–534.

[34] P. Shah, B. C. S. Reddy, S. N. Merchant, and U. B. Desai
"Context enhancement to reveal a camouflaged target and to assist target localization by fusion of multispectral surveillance videos,"
Signal Image Video Process., vol. 7, no. 3, pp. 537–552, 2013.

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson
"How transferable are features in deep neural networks?" in
Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 3320–3328.

[36] K. Soomro, A. R. Zamir, and M. Shah
"UCF101: a dataset of 101 human action classes from videos in the wild.," CRCV-TR-12-01, November, 2012.

[37] C. F. Hester and D. Casasent
"Multivariant technique for multiclass pattern recognition,"
Appl. Opt., vol. 19, no. 11, pp. 1758–1761, 1980.

[38] A. Ulhaq, X. Yin, Y. Zhang, and I. Gondal
"Action-02MCF: a robust space–time correlation filter for action recognition in clutter and adverse lighting conditions," in
Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst. Berlin, Germany: Springer, 2016, pp. 465–476.

[39] J. Lewis, S. Nikolov, A. Loza, E. F. Canga, N. Cvejic, J. Li, A. Cardinali, C. Canagarajah, D. Bull, T. Riley, D. Hickman,

and M. I. Smith
"The Eden Project multisensory data set," The Online Resource for Research in Image Fusion (ImageFusion.org), 2006.

[40] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann
"InfAR dataset: infrared action recognition at different times,"
Neurocomputing, vol. 212, pp. 36–47, 2016.

[41] B. V. K. V. Kumar, A. Mahalanobis, and R. D. Juday
Correlation Pattern Recognition. New York, NY, USA: Cambridge University Press, 2005.

Anwaar Ulhaq is serving as a lecturer in School of Computing and Mathematics and deputy leader in Machine Vision and Digital Health Research Group at Charles Sturt University, New South Wales, Australia. Anwaar holds PhD (Artificial Intelligence) from Monash University, Australia. He has also worked a research fellow at Institute for Sustainable Industries & Liveable Cities, Victoria University, Australia. His research interests include signal and image processing, deep learning, data analytics and computer vision.