Action Recognition in the Dark via Deep Representation Learning

Anwaar Ulhaq
School of Computing and Mathematics,
Charles Sturt University, Port Macquarie, NSW, Australia, 2444.
Email: aulhaq@csu.edu.au

***Abstract*—Human action recognition for automated video surveillance applications is an interesting but a daunting task especially if the videos are captured in unfavourable lighting conditions. These situations encourage the use of multi-sensor video streams. However, simultaneous activity recognition from multiple video streams is a difficult problem due to their complementary and noisy nature. This paper proposes simultaneous action recognition from multiple video streams using deep multi-view representation learning. Furthermore, it introduces a spatio-temporal feature based correlation filter, for simultaneous detection and recognition of multiple human actions in low-light conditions. We evaluated the performance of our proposed filter with extensive experimentation on night-time action datasets. Experimental results indicate the effectiveness of deep fusion scheme for robust action recognition in extremely low-light conditions.**

I. Introduction

Robust recognition of human actions in diverse scenarios and lighting conditions constitutes a challenging research problem in computer vision. Various potential applications in diverse areas (e.g. visual surveillance, video retrieval, sports video analysis, human computer interfaces, and smart rooms) have encouraged the development of a large number of automated action recognition techniques. These approaches consider many challenging scenarios like actions in the wild (YouTube videos) [1], actions in the crowd [2],actions in group formations [3], actions in movies [4], actions across different viewpoints [5] and in the presence of occlusion [6].

However, the major approaches consider action recognition in high quality day-time video sequences or in the presence of bright lighting conditions and do not focus on adverse lighting or the recognition of actions during night-time. We consider the case of low-light scenes in an environment with insufficient illumination. Any visual processing and manipulation of intensity values in such imagery results in different types of undesired artefacts like noise amplification, intensity saturation, and loss of resolution. This problem motivates the use of multiple sensors often of complementary nature. A general multi-sensor night vision system employs low light images by low light visible cameras and infra-red images by forward looking infra-red cameras. Detecting and recognizing any human action from these videos is a daunting task for any human observer or operator. Fig 1 illustrates these scenarios.

Fig. 1. Two representative scenarios of night-time imagery captured by low light visible CCD (A) sensor and infra-red sensors (B). These images presents visual information of complementary nature and lack certain visual information on individual basis. Such kind of degraded context provides significant challenge for action recognition.

Multi-view representation learning [7] has getting increasing attention of machine learning research community. Multi-views data contains complementary information, and multi-view learning can exploit it to learn representation that is more discriminative than that of single-view learning methods. Canonical correlation analysis (CCA) is the fundamental work which other extensions are based on. Kernal trick was introduced by [8] which find non-linear projections of two views. Another extension [9] is generalize it by considering more views of data.

Our work is based upon the idea that multi-views (multi-spectral) data is valuable as it contains complementary and common information, so multi-view learning can exploit it to learn multi-spectral discriminative representations better than any single-view learning strategy. In this work, we introduce a deep fusion framework for robust action recognition across different visual spectrums. At first, spatio-

978-1-7281-0247-4/18/$31.00 ©2018 IEEE

temporal convolutional features are learnt from each video stream. We integrate important visual cues from these features using deep multi-view representation learning. We then, train feature based correlation filter to recognize actions. The motivation of using correlation filter is simultaneous detection and recognition of multiple actions in a video stream. We claim the following contributions in this paper:

1- We consider the problem of recognizing human actions form low quality video sequences captured in low lighting conditions and multi-sensor scenario. It is in contrast to the majority of available action recognition work that focuses on high quality and bright imagery. This investigation is beneficial for developing many new application of action recognition like video monitoring and surveillance in unfavourable lighting conditions.

2- To the best of our knowledge, our proposed framework presents a novel approach to fuse deep correlations from multiple stream using deep multi-view representation learning. To speed up, simultaneous action detection and recognition, we use correlation filter.

The organization of the paper can be described as: In next subsection, we discuss the proposed framework II. Experimental results are presented in section III followed by conclusion and references.

A. Prior Work

Over the last decade, various approaches have been proposed for human action recognition that vary in terms of used technique, gained performance and reduced complexity. The proposed techniques for human action recognition can be categorized on the basis of *representation* used for encoding an action. Notable research work include the extraction of spatio-temporal features [10], [11], space-time templates [12]–[14], trajectories-based descriptors [15] and action filters [16], [17]. Different surveys [18], [19] provide recent development on action recognition approaches.

Correlation filters:

Correlation filters [20]–[22] have recently received significant attention. These filters overcome shortcomings of spatial template matching. The success of correlation filters in tracking applications has inspired their usage in action recognition. A representative work is Action MACH filter [16]. The performance of Action MACH is encoring. However, the capability of Action MACH, has been questioned by [23]. A multi-class correlation filter is proposed in [24]. Most recently, a multi-class correlation filter framework [**?**] is proposed for cross-view action recognition. However, all above approach assume good quality video sequences.

Action recognition in low-quality videos:

Action recognition in low-quality night-time videos is comparatively less explored. Thermal infra-red imagery is used for human action recognition in [25] . The algorithm uses well-known features like histogram of oriented gradients and nearest neighbour classification. Another work [26] uses gait energy images. The recognition performance is low and restricted only to walking activity which is easier to recognize. Above approach are very similar to action recognition of daytime videos as they use silhouette images. Another work proposes contextual action recognition based on 3D FFT and contextual cues is proposed in [27]. This approach uses context of night vision into consideration but its recognition performance is very low.

This paper proposes a robust action recognition approach to deal with extremely low quality night-time video sequences. Our framework is based on deep convolutions, frequency domain correlation analysis and can detect and classify human actions in a robust manner.

II. The Proposed Framework

In this section, we present and discuss our proposed framework for action recognition based on deeply fused convolutional features. We first use spatio-temporal features, C3D features [28] as building blocks of our framework. The motivation of using C3D features is their recent success in human action recognition [28] and availability of trained action models. We then use it to extract the proposed deep discriminative features. We then describe design of action correlation filter for recognition of action classes. An illustration of our framework is presented in fig. 2.

Notation: We will use the following notations: we will use small letters for column vectors, capital letters for matrices and bar (-) for frequency domain representation.

A. C3D : Spatio-Temporal Features

The 2D Convolutional neural networks can only analyze and learn the spatial features present in the images. In order to learn spatio-temporal features, we use C3D features [28]. C3D network is based on 3D convolutions and pooling to retain the time variant information in the video. The network following the footsteps of [28] uses a 16 frames clip each of size 112×112 as an input. It contains 8 convolution layers, 5 pooling layers, 2 fully connected layers followed by a softmax classification layer. Each of the convolution layers has ReLU activation applied over it with padding to preserve the spatial dimensions. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. All the parameters of 3D-CNN architecture are learned discriminatingly to produce a single compact descriptor for action recognition. For each video clip, fc6 activations are averaged to form a 4096-dimensional video descriptor which is then followed by an L2-normalization. This video descriptor/feature is C3D. For a given class of actions taken from n number of video streams, we extract C3D features for each stream separately. In our case $n = 2$.

B. Deeply fused Convolutional Features

The discriminative strength of C3D features [28] is good fit only for single spectrum as these features are single channel representations. Our goal, however is to learn fused representation from multiple video streams at a given time.

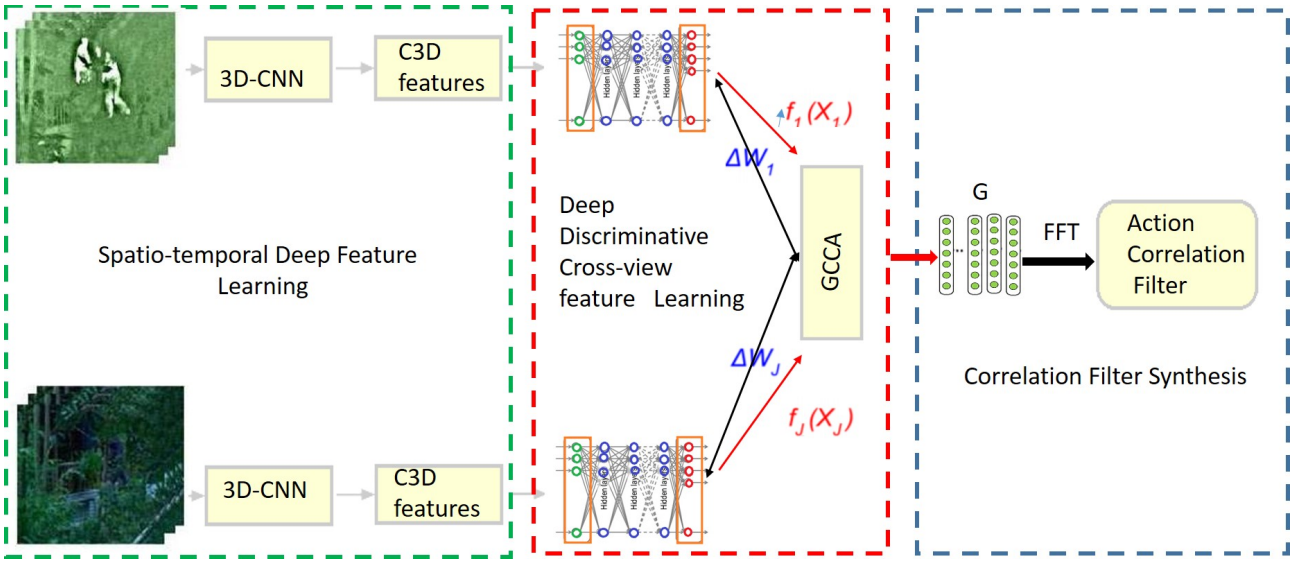

Fig. 2. An illustration of the proposed framework: Actions are represented to CCN networks for extract C3D features, In Stage 2, these features are used as input to multi-view deep networks to learn deeply fused convolutional features using deep version of GCCA (Generalized canonical correlation analysis). These features are then used to train action correlation filter to find the action class.

The idea of deeply fused convolutional features is built upon the concept of multi-view representation learning [**?**]. We use a multi-view deep network which extracts representations of different spectrums by presenting them to multiple stacked layers of nonlinear transformation and back-propagating the gradient of the objective to tune each stream's network. It does so by learning a nonlinear map for each video stream in order to maximize the correlation between the learnt model across views. The overall objective is to train networks that reduce the overall reconstruction error among their outputs.

Let N be the number of extracted C3D features from video with actions captured from a single viewpoint, J the number of available viewpoints, $X_j \in \mathbb{R}^{dj \times N}$ the input feature matrix for j^{th} video stream, m the number of layers in each network, n the number of neurons in each layer and $f_j(X_j)$ represents the j^{th} network outputs. The output of the m^{th} layer for the j^{th} stream is $o_m^j = S(W_m^j o_{m-1}^j) + b_m^j$ with S being the Sigmoid function (non-linear activation), $W_m^j \in \mathbb{R}^{n_m \times n_{m-1}}$ is the weight matrix for for m^{th} layer and for j^{th} view network.

The overall deep fusion objective can be written as an optimization problem. If d_j is the dimensionality of the j^{th} view, k is the dimensionality of the learned representation, L_j as linear (of output of j^{th} network output) transformations, the objective is to find the weight matrices $W^j = \{W_1^j \dots W_m^j\}$ as follows:

$$\begin{aligned} \underset{G \in \mathbb{R}^{k \times N}, L_j \in \mathbb{R}^{j \times k}}{\text{minimize}} \quad & \sum_{j=1}^{J} ||G - L_j f_j(X_j)||_F^2 \\ \text{subject to} \quad & GG^T - I_k. \end{aligned} \tag{1}$$

where $||.||_F^2$ is the matrix Frobenius norm, $L_j f_j(X_j)$ is view-specific representation and $G \in \mathbb{R}^{k \times N}$ is the stream-independent shared representation. This optimization problem is solved using stochastic gradient descent (SGD) with mini-batches.

The above problem can be solved by solving an eigenvalue problem. If covariance matrix of j^{th} network output is expressed as $C_{jj} = f(X_j)f(X_j)^T$, then $P_j = f(X_j)^T C_{jj}^{-1} f(X_j)$ is the corresponding projection matrix. It is a linear transformation that transforms data with a known C_{jj} into a set of new variables whose covariance is the identity matrix, meaning that they are uncorrelated), symmetric and positive semi-definite matrix. For J view, we have $M = \sum_{i=1}^{J} P_j$ and we can write G (a shared representation) where rows of G are top k singular values of M and $L_j = C_{jj}^{-1} f(X_j) G^T$, the minimum of objective function in equation 1 can be re-written as:

$$\begin{aligned} \sum_{j=1}^{J} ||G - L_j f_j(X_j)||_F^2 = \sum_{j=1}^{J} ||G - G f_j(X_j) C_{jj}^{-1} f_j(X_j)||_F^2 \\ = kJ - Tr(GMG^T). \end{aligned} \tag{2}$$

Minimization of the objective function in equation 1 is equivalent to maximization of $Tr(GMG^T)$, which is the sum of eigenvalues $Y = \sum_{i=1}^{k} \lambda_i(M)$. By taking derivative of Y w.r.t each network output $f_j(X_j)$, we get:

$$\frac{\partial Y}{\partial f_j(X_j)} = 2L_j G - 2L_j L_j^T f_j(X_j) \tag{3}$$

It indicates the radiant as the difference between k-dimensional representation G and the projection of the original data. Intuitively, if this difference is large, the network weight should reverie a large update.

We rephrase that $L_j f_j(X_j)$ is video streeam-specific representation and $G \in \mathbb{R}^{k \times N}$ is the spectrum-independent shared representation in which we are interested to learn. For N C3D features collected from videos taken from j video stream, we learn G representations and call them deeply fused convolutional features. These features are spectrum-independent representation and can be use to train feature-based correlation filter for a given action class.

Algorithm 1 deeply fused convolutional feature Learning Algorithm

for $a \leftarrow 1, nClass$ **do**
  for $j \leftarrow 1, nViews$ **do**
    $[F_1, F_2, \dots F_j] \leftarrow$ calculate $C3D$ features
  end for
end for
for $a \leftarrow 1, nClass$ **do**
initialize weights $[W_1, W_2, \dots W_j]$
  for $i \leftarrow 1, nIterations$ **do**
    for $j \leftarrow 1, nViews$ **do**
      $o_j \leftarrow$ forward pass of F_j with W_j
      mean-center o_j
    end for
    $L_1, L_2, \dots L_j, G_a \leftarrow gcca(o_1, o_2, \dots o_j)$
    for $j \leftarrow 1, nViews$ **do**
      $\partial Y/\partial o_j \leftarrow L_j L_j^T o_j - L_j G_a$
      $\nabla W_j \leftarrow backpropagate((\partial Y/\partial o_j, W_j)$
      $W_j \leftarrow W_j - \alpha \nabla W_j$
    end for
  end for
  for $j \leftarrow 1, nViews$ **do**
    $o_j \leftarrow$ forward pass of F_j with W_j
    mean-center o_j
  end for
  $G_a \leftarrow gcca(o_1, o_2, \dots o_j)$
end for

C. Action Correlation Filter

The motivation of using correlation filters is their flexibility and speedup due to working in frequency domain. The filter design problem is posed as an optimization problem.Support vector machines (SVMs) are robust classifiers due to their maximizing margin property. On the other hand, correlation filters (CFs) are good at localization property. Therefore, by combining both properties, a better filter can be synthesized.

Given N deep multi-view representative features, a multi-objective function of this filter is given as:

$$\min_{h,b}(h^T h + \varphi \sum_{i=1}^{N} \xi_i, \sum_{i=1}^{N} ||h \otimes f_i - g_i||_2^2), \quad s.t. y_i(h^T . f_i + b) \geqslant c_i - \xi_i \tag{4}$$

where h denotes filter, ξ_i is the penalty term, φ trade-off parameter, $c_i = 1$ for true-class samples and $c_i = 0$ or other small value for false-class samples and $h^T h + \varphi \sum_{i=1}^{N} \xi_i$ is the margin criterion and $\sum_{i=1}^{N} ||h \otimes f_i - g_i||_2^2$ is the localization criterion.

For better localization, we desire a peak centered at the targets location and zeros everywhere else. Therefore, we use, $g_i = \{0, - - - -, 0, h^T f_i, 0, - - - - 0\}^T$ and the cross correlation value of f_i and h is considered as target location. The above objective function can be reduced to a quadratic function:

$$\min_{\bar{h}, \acute{b}} \bar{h}^\dagger Z \bar{h} + \acute{\varphi} \sum_{i=1}^{N} \xi_i, \quad s.t. y_i(\bar{h}^\dagger \bar{f}_i + \acute{b}) \geqslant c_i - \xi_i \tag{5}$$

where $Z = \lambda \mathcal{I} + (1 - \lambda) D$ with $\mathcal{I}$ is the identity matrix and cross power spectrum matrix $\vec{D}$ is $D = \sum_{i=1}^{N} \bar{F}_i^{1\dagger} \vec{F}_i^1$, $\acute{b} = b \times d$ where d is the dimensionality of f_i and $\bar{F}_i$ is a diagonal matrix whose diagonal entries are the elements of $\bar{x}_i$.

if we put $\bar{h} = \bar{Z}^{\frac{1}{2}} \bar{h}$ and $\bar{f}_i = \bar{Z}^{-\frac{1}{2}} \bar{f}_i$. Since $\bar{Z}$ is positive definite matrix we can write the objective function as:

$$\min_{\bar{h}, \acute{b}} \bar{h}^\dagger \bar{h} + \acute{\varphi} \sum_{i=1}^{N} \xi_i, \quad s.t. y_i(\bar{h}^\dagger \bar{f}_i + \acute{b}) \geqslant c_i - \xi_i \tag{6}$$

It becomes an SVM objective function and can be implemented using a standard SVM solver. This correlation filter can simultaneously localize and classify actions of interest as it integrates rich discriminations of deep feature space, generalization of maximum margin classification criteria, better localization performance of correlation filtering into a single classifier. We named our filer as 3D-SDCF.

III. Experimental Results and Discussion

The performance of the proposed filter is evaluated using a comprehensive set of experiments on challenging human action data sets in both day-time and night-time scenarios. This section provides description of used datasets, details of experimental set-up, action recognition and detection accuracy with comparative performance evaluation.

A. Dataset and Experimental set-up

Our experimentation is based on Night-Vision Video Data set (NV) [24]. This dataset represent a good collection of human actions performed in daytime and night-time settings with different set of challenges for recognizing actions. We divide all datasets into training and testing videos and use 70% videos for training and 30% for testing.

Night Vision Action Dataset (NV) [24]: This dataset contains video sequences recorded by using two separate cameras: Raytheon Thermal IR-2000B and Panasonic WV-CP470. The registration of visual and IR videos is performed by manually selecting correspondences in both corresponding frames, followed by computing a least-squared

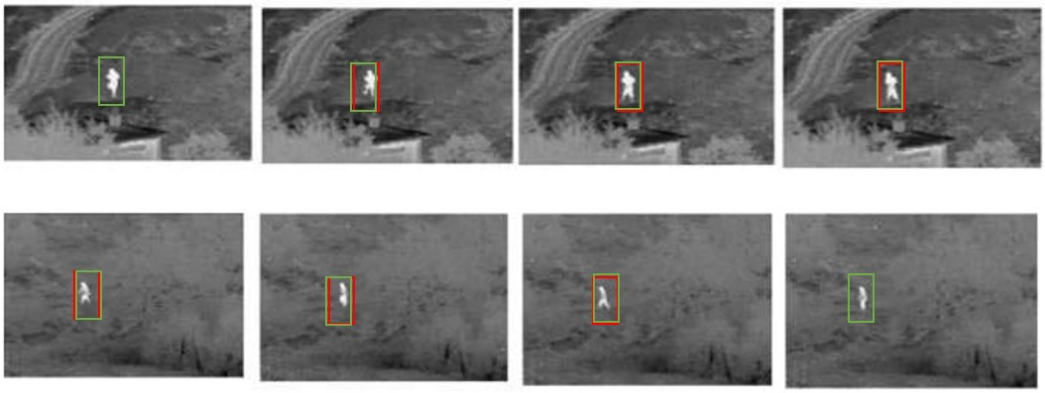

Fig. 3. Action instance detection in three NV Actions instances where red bounding box is the actual ground truth while green bounding box shows detection by 3D-SDCF.

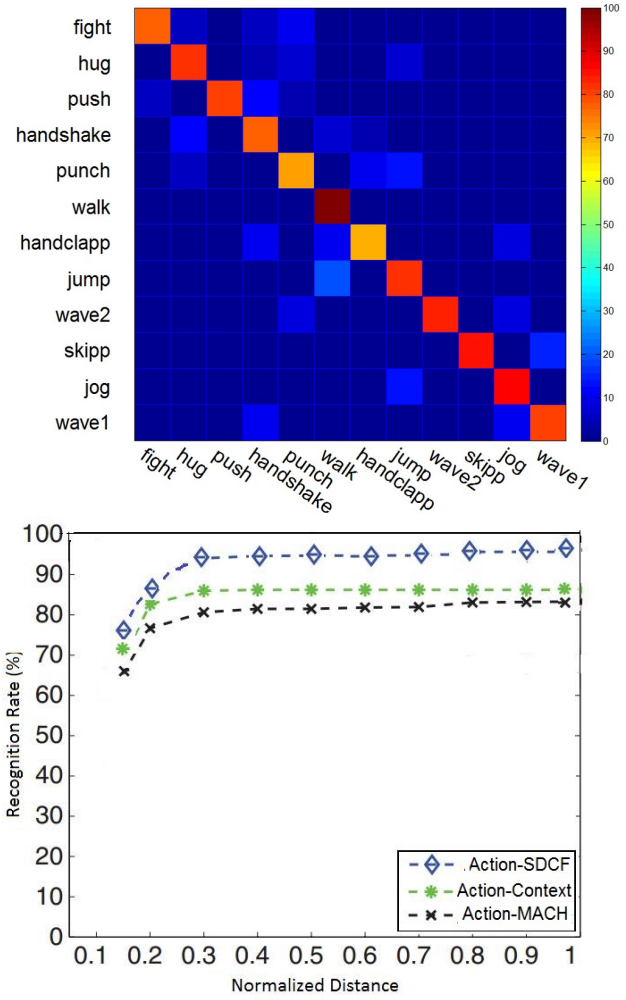

Fig. 4. (Top) Confusion matrix showing for NV action dataset.(Bottom) Plot of recognition rate vs normalized correlation error (E) for different approaches. It indicates that for every value of E, 3D-SDCF outperforms other classifiers.

error by fitting homography for each stream. IR video frames were warped to align with the visual pixels. In addition to these videos, this dataset includes another 20 video sequences from TNO image fusion dataset, visible-infrared sequences and Ohio-state University thermal dataset. The video sequences contain ten action categories of walking, wave1, wave2, stand-up, sit-down, clapping, pick-up, punching and running actions performed by multiple actors.

Each video is split into non-overlapped 16-frame clips and then given as input to the deep network. We use 70-30 ration of training and evaluation set. For each action category, we synthesize a different 3D-SDCF filter and correlate the synthesized filter with test video in search of correlation peak. These filters are trained on extracted spatio-temporal convolutional feature maps from original video sequences We calculate confusion matrices for NV datasets and display it in fig. 3 . Due to availability of single sensor videos, we used single stream 3D-SDCF for these datasets.

Similarly, we observed that our filter performance for NV action dataset outperforms other competitive approaches . We achieved an overall average precision of 80.3%. We achieved low performance with average precision of 75.0% and 71.1% for punch and handclasp actions. It was due to marginal motion in both action instances. This performance is still better to other competitive approaches. It is comparable to other state-of-the art approaches like Dense-Traj [15], MBH [29] and [30]. A detailed comparison is summarized in TABLE 1 (right).

Another experiment was performed to compare no fusion and deep fusion performance evaluation. Single stream framework used single channel or already fused video sequences. In night vision scenario, often infra-red and visible sequences are fused for context enhancement and better situational awareness. Different fusion approaches depend on the data of data availability. We use single stream filter training illustrated in figure for such data. NV action dataset was more suitable for this experimentation. In contrast , in two stream training network , we fuse feature

Method	Average precision
STIP [10]	49.1%
3D SIFT [31]	49.5%
HOF [4]	68.5%
Dense-Traj [15]	68.66%
MBH [29]	67.3%
CNN Features [30]	76.6%
3D- SDCF	80.3%

TABLE I
COMPARATIVE ANALYSIS IN TERMS OF AVERAGE PRECISION FOR NV ACTION DATASET.

maps in the later stage at layer 5 level and feed fused feature map to the filter.

IV. CONCLUSION

This paper proposes a novel approach for action recognition in extremely low light condition by deeply fusion information from multiple video streams. Deep multi-view learning is used for fusing convolutional features. Spatio-temporal correlation filter is used for recognizing and detecting human actions in adverse lighting conditions and clutter. The proposed filter takes advantage of the maximum margin property of SVMs with localization capability of correlation filtering and meaningfulness of deep spatio-temporal convolutions into a correlation filter framework. When we evaluate it against variety of human action datasets captured under different challenging condition, we found that such kind of synergy contributes towards enhanced recognition performance.

REFERENCES

[1] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *IEEE Conference on Computer Vision and Pattern Recognition,*, pp. 1996–2003, 2009.

[2] P. Siva and T. Xiang, "Action detection in crowd.," in *British Machine Vision Conference,*, pp. 1–11, 2010.

[3] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *IEEE International Conference on Computer Vision,*, pp. 742–749, 2003.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[5] A. Ulhaq, X. S. Yin, J. He, and Y. Zhang, "On space-time filtering framework for matching human actions across different viewpoints," *IEEE Transactions on Image Processing* **27**(3), pp. 1230–1242, 2018.

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

[7] Y. Li, M. Yang, and Z. Zhang, "Multi-view representation learning: A survey from shallow methods to deep methods," *CoRR* **abs/1610.01206**, 2016.

[8] D. Hardoon, J. Mourão-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fmri data using kernel canonical correlation," **37**, pp. 1250–9, 11 2007.

[9] P. Horst, "Generalized canonical correlations and their applications to experimental data," *Journal of Clinical Psychology* **17**(4), pp. 331–347.

[10] I. Laptev, "On space-time interest points," *International Journal of Computer Vision* **64**(2-3), pp. 107–123, 2005.

[11] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65–72, IEEE, 2005.

[12] A. Bobick and J. Davis, "he recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(3), pp. 257–267, 2001.

[13] B. Yao and S.-C. Zhu, "Learning deformable action templates from cluttered videos," in *12th International Conference on Computer Vision*, pp. 1507–1514, IEEE, 2009.

[14] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *International Conference on Computer Vision*, **2**, pp. 1395–1402, 2005.

[15] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision* **103**(1), pp. 60–79, 2013.

[16] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, June 2008.

[17] G. I. Haq Anwaar and M. Manzur, "Action recognition using spatio-temporal distance classifier correlation filter," in *International Conference on Digital Image Computing Techniques and Applications (DICTA),*, pp. 474–479, 2011.

[18] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing* **60**, pp. 4–21, 2017.

[19] M. Vrigkas, C. Nikou, and I. Kakadiaris, "A review of human activity recognition methods," *Front. Robot. AI 2: 28. doi: 10.3389/frobt* , 2015.

[20] J. A. Fernandez, V. N. Boddeti, A. Rodriguez, and B. V. K. V. Kumar, "Zero-aliasing correlation filters for object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **37**(8), pp. 1702–1715, 2015.

[21] V. N. Boddeti and B. Kumar, "Maximum margin vector correlation filter," *arXiv preprint arXiv* , p. 1404.6031, 2014.

[22] V. N. Boddeti, T. Kanade, and B. V. K. V. Kumar, "Correlation filters for object alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition,*, pp. 2291–2298, 2013.

[23] S. Ali and S. Lucey, "Are correlation filters useful for human action recognition?," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 2608–2611, Aug 2010.

[24] 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 *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 465–476, Springer, 2016.

[25] H. Eum, J. Lee, C. Yoon, and M. Park, "Human action recognition for night vision using temporal templates with infrared thermal camera," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2013 10th International Conference on*, pp. 617–621, IEEE, 2013.

[26] J. F. Li and W. G. Gong, "Application of thermal infrared imagery in human action recognition," in *Advanced Materials Research*, **121**, pp. 368–372, Trans Tech Publ, 2010.

[27] H. Anwaar, G. Iqbal, and M. Murshed, "Contextual action recognition in multi-sensor nighttime video sequences," in *Digital Image Computing Techniques and Applications (DICTA),*, pp. 256–261, 2011.

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

[29] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*, pp. 428–441, Springer, 2006.

[30] 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* **212**, pp. 36–47, 2016.

[31] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360, ACM, 2007.