Pattern Recognition 122 (2021) 108282

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Fine-grained action recognition using dynamic kernels

Sravani Yenduri*, Nazil Perveen, Vishnu Chalavadi, Krishna Mohan C

Indian Institute of Technology Hyderabad, Kandi, Telangana 502285, India

ARTICLE INFO

Article history:
Received 16 April 2021
Revised 20 August 2021
Accepted 28 August 2021
Available online 30 August 2021

Keywords:
Fine-grained action recognition
Spatio-temporal features
Gaussian mixture model
Dynamic kernels

ABSTRACT

Fine-grained action recognition involves comparison of similar actions of variable-length size consisting of subtle interactions between human and specific objects. Hence, we propose a dynamic kernel-based approach to handle the variable-length patterns for effective recognition of fine-grained actions. Initially, we extract local spatio-temporal features for each video to capture appearance and motion information effectively. An action-independent Gaussian mixture model (AIGMM) is trained on the extracted features of all fine-grained actions to analyze spatio-temporal information and preserve the local similarities among fine-grained actions. Then, the statistics of AIGMM, namely, mean, covariance, and posteriors are used to build the kernels for finding the similarity between any two fine-grained actions by mapping statistics to kernel feature space. We demonstrate the effectiveness of proposed approach using three dynamic kernels i.e., GMM mean interval kernel, supervector kernel, intermediate matching kernel on four varieties of fine-grained action datasets, namely, MERL, JIGSAWS, KSCGR, and MPII cooking2

1. Introduction

The task of action recognition is to classify an action present in a video from a set of known actions. The actions can be broadly categorized into two types, namely, coarse-grained actions and fine-grained actions. The coarse-grained actions involve full-body activities, especially *jumping, diving, cooking*, etc., which are differentiated effortlessly due to high intra-class similarity. Whereas, the fine-grained actions such as *take plate, put in cupboard, wash objects*, etc., involving subtle interactions between human and objects are difficult to distinguish because of the presence of diverse objects, high inter-class, and low intra-class similarities. Fine-grained action recognition is used in many applications like human-computer interaction, robotics, surveillance [1,2], video description [3], and autonomous vehicles because of its ability to discriminate the visually similar actions as shown in Fig. 1. Fine-grained action recognition is a challenging task due to occlusion of objects or actions, different duration in performing the same action, view-point variations, etc.

Fine-grained action recognition requires extracting efficient spatio-temporal features that provides both spatial and temporal cues. Traditional hand-crafted approaches, namely, spatio-temporal interest points (STIP) [5], 3D SIFT [6], and improved dense trajectories (IDT) [7] extracts the spatio-temporal features to provide a video-based representation for action recognition. However, the number of feature vectors varies from one video clip to another due to (i) variable number of interest points or sampled feature points obtained during the extraction of STIP or IDT features, and (ii) the variation in duration of a fine-grained action from one actor to the other. For example, the duration of a fine-grained action '*reach to shelf*' is 68 frames in Fig. 2a, whereas the same fine-grained action takes 87 frames in Fig. 2b. Conventionally, Gaussian mixture models (GMMs) or hidden markov models (HMMs) are employed to aggregate these variable-length features by estimating the parameters of GMM or HMM to classify the fine-grained actions [8]. Later, deep learning approaches such as multi-stream networks [9], 3D-convolutional neural networks (3D-CNNs) [10] are explored to obtain the spatio-temporal representation. These networks handle the variable-length features by combining the feature vectors either by pooling or employing long short term memory (LSTM) or Bi-directional long short term memory (BiLSTMs) [11]. However, these networks require computation of large number of parameters, and are hard to train from end-to-end [12].

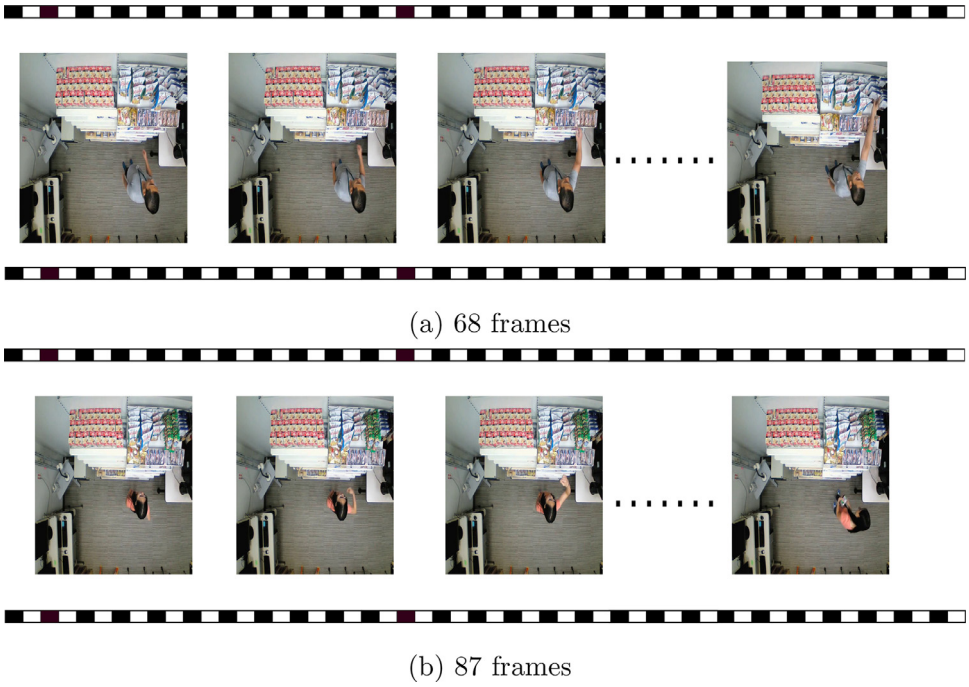
To overcome the above challenges, we propose an approach for fine-grained action recognition using dynamic kernels. Initially, we extract spatio-temporal features, namely, histogram of optical flow (HOF) and motion boundary histogram (MBH) for each video clip. A large GMM, known as action independent GMM (AIGMM) is built on the extracted features of all classes to model the subtle variations among fine-grained actions. Here, AIGMM aims to capture the attributes representing fine-grained actions. Attributes are the basic units that collectively form a fine-grained action. For ex-

* Corresponding author.
E-mail addresses: cs18resch02001@iith.ac.in (S. Yenduri), cs14resch11006@iith.ac.in (N. Perveen), cs16m18p000001@iith.ac.in (V. Chalavadi), ckm@cse.iith.ac.in (K.M. C).

https://doi.org/10.1016/j.patcog.2021.108282
0031-3203/

(a) cut (b) peel

Fig. 1. Illustration of fine-grained actions in cooking activity from MPII cooking2 dataset [4]. An example of two visually similar actions (a) **cut** and (b) **peel**.



(a) 68 frames

(b) 87 frames

Fig. 2. Duration of '*reach to shelf*' fine-grained action from MERL dataset [13].

ample, a *cut* action consists of attributes such as right-hand retract, left-wrist rotate etc. The parameters of AIGMM such as mean, covariance, and posteriors are used to find the similarity among fine-grained actions by adopting various kernel methods. The kernel methods calculate the distance between the local feature vectors & the parameters of AIGMM and transform the calculated distances to high-dimensional feature space for better discriminability [14]. The performance of kernel methods depends on the choice of kernel function. The kernel functions that focuses on handling the varying length feature vectors are referred as dynamic kernels. Dynamic kernels handle the variable-length features by either mapping to patterns of fixed length (probability based kernels) or selecting the best feature vectors (matching based kernels). The probability based dynamic kernel utilizes the first-order (mean) & second-order (covariance) statistics for calculating the distances. Whereas, the matching based kernels select the local feature vectors that are close to AIGMM statistics for constructing the kernel. We explore various dynamic kernels to classify the fine-grained actions which are of varying length. The effectiveness of our approach is demonstrated on 4 varieties of challenging fine-grained action datasets, namely, MERL, JIGSAWS, KSCGR, and MPII cooking2. The main contributions of this paper can be summarized as:

- We construct an action-independent GMM (AIGMM) using the local spatio-temporal features to preserve the local similarity among the fine-grained actions.
- We propose an approach to handle the variable-length patterns of fine-grained actions by mapping the statistics of trained AIGMM onto kernel feature space.
- Explored various dynamic kernels on 4 varieties of challenging fine-grained action datasets, namely, MERL, JIGSAWS, KSCGR, and MPII cooking2. The fine-grained actions in these datasets exhibits the issues like high intra-class variability, low inter-class variability, and occlusion.

2. Related work

In this section, we discuss various existing approaches in the literature for both coarse-grained and fine-grained action recognition tasks. We also discuss different dynamic kernels used in various domains to encode the variable-length patterns in this section.

2.1. Coarse-grained action recognition

Traditional methods represent the actions in a video by using several features, namely, spatio-temporal interest points (STIP) [5], IDT [7], etc. The STIP utilizes the Harris 3D corner detector to obtain the interest points that are tracked throughout the video. In order to capture both spatial and temporal information, histogram of oriented gradients (HOG) and histogram of optical flow (HOF) features are extracted around these detected points. Similarly, the densely sampled feature points are tracked in IDT across several frames to obtain the trajectory information. The HOG, HOF, and motion boundary histogram (MBH) features extracted around these feature points can effectively represent the spatio-temporal cues of an action. To encode the above-mentioned features, several aggregation frameworks such as BoW [15], VLAD [16], Fisher vector [17], etc. are analyzed by clustering these features using either k-means algorithm or Gaussian mixture models (GMMs). These frameworks exploit the multiple statistics of the clusters for better discrimination of actions.

Li et al. [18] extended the VLAD [16] by encoding the deep features for efficient action representation. To overcome the absence of temporal information in conventional CNNs, Tu et al. [9] proposed a multi-stream CNN to recognize the human actions by considering the most discriminative regions that are obtained using a motion saliency measure. It consists of two independent streams to capture the appearance and motion information of the regions of interest. Moreover, the multi-stream networks train the temporal and spatial streams independently. Wangli et al. [19] explored various strategies to incorporate interaction between the temporal and spatial streams. It is claimed that the dense connections between these two networks integrate the spatial and temporal information at the feature representation level and knowledge distillation at higher-level layers. However, the multi-stream networks [20,21] do not effectively model long-term temporal information due to the limited temporal window size i.e., single frame for spatial stream and a stack of few frames for temporal stream. Hence, Wang et al. [22] introduced a temporal segment networks (TSN) to model long-term temporal structure by adopting a novel temporal sampling strategy to obtain the video-level representation for each action.

Later, 3D-CNNs [10,23] are introduced to capture the spatio-temporal information effectively by introducing 3D convolution and mixed convolution filters, respectively. In order to address the issue of computational overhead [12] of these models, Yang et al. [24] proposed an asymmetric 3D convolutional network which helps in the reducing the number of parameters and hence the lower computational complexity. However, the fixed structure of the 3D convolution filter in both spatial and temporal dimensions restrict the learning capacity of action representation. Jun et al. [25] addressed this problem by introducing the deformable attentive spatio-temporal 3D convnets to capture the appearance and irregular motions of action efficiently. Similarly, temporal shift module (TSM) [26] is incorporated into 2D CNNs to model temporal information without additional computational cost. Although these approaches can effectively classify coarse-grained actions such as *lifting, diving*, and *running*, etc., they fail to model subtle interactions between the human and objects, which are crucial for fine-grained action recognition.

2.2. Fine-grained action recognition

The fundamental challenge in fine-grained action recognition is to recognise the actions that are visually similar to each other and have subtle variations in motion. The global pooling of low-level features restricts the local interaction motion information, thus attenuating the prominent discriminative information for fine-grained action recognition [27]. Zhou et al. [27] presented a mid-level approach by constructing the sub-graphs for recognising these subtle variations. Initially, a spatio-temporal graph is constructed in which nodes are connected based on the appearance similarity and trajectory strength. Here, nodes are the interaction regions generated by BING [28]. This graph is divided into sub-graphs using a graph segmentation algorithm. These sub-graphs are known to contain the information of interactions between human and objects. Similarly, a graph is built with nodes as the object proposals that capture the discriminative features of fine-grained actions [29]. Here, the object proposals are obtained by merging the interaction regions extracted in conjunction of the saliency map representing the histogram of motion.

Later, Miao et al. [30] proposed a six-stream region-based CNN to address the problems of coarse-grained and fine-grained action recognition.The framework consists of 6 independent streams whose input images contain both appearance and motion cues by cropping them at different scales. The frames are cropped to human region and interaction region to obtain the global and local information of a fine-grained action. The feature descriptors from 6 streams are concatenated to form an efficient representation for better discrimination of fine-grained actions. However, this approach considers the spatial regions to contain prominent information of fine-grained actions neglecting the motion information during interactions. To capture the spatio-temporal information, additional blocks are incorporated to attend to clues crucial for fine-grained action recognition [31,32]. Recently, Han et al. [32] utilized the triplet loss for training the convolutional neural network [33] to reduce the intra-class variance and increase the inter-class distance.

2.3. Dynamic kernels

Dynamic kernel based approaches are explored in literature to handle the varying length data such as speech [34,35], image [36]. One of the dynamic kernels i.e., GMM supervector kernel based SVM [37] is constructed by training a standard GMM using the maximum a posterior (MAP) adaptation to improve the classification performance. The adapted means of GMM are stacked to form a mean supervector in order to deal with the view-point and actor variations. Chang et al. [38] introduced the mean interval kernel (MIK) by extending the Bhattacharyya based SVM kernel for better discrimination. The MIK exploits the first order and second-order statistics of GMM to capture the underlying useful information.

Boughorbel et al. [36] introduced a computationally efficient dynamic kernel i.e., intermediate matching kernel (IMK) for object recognition. An IMK matches the set of local features by constructing a set of virtual features. These virtual features play the role of feature selectors by choosing the closest local features based on GMM mixtures. As the number of virtual features is less than the local features, the computation time for IMK is low.

Several methods have been explored to obtain the discriminative representation for coarse-grained and fine-grained actions. However, the major limitations of these approaches are: (i) inability to generalize well on the smaller datasets as they require a large amount of labelled data for training, (ii) they are computationally expensive, (iii) these networks obtain a representation by employing the reduced version of the original frame. This might cause a loss of contextual information which limits the networks to achieve better discrimination, and (iv) most of the approaches opt for sampling strategies to obtain the fixed input size. In contrast, our approach does not involve any kind of sampling strategy and hence it preserves the subtle interactions among fine-grained actions effectively.

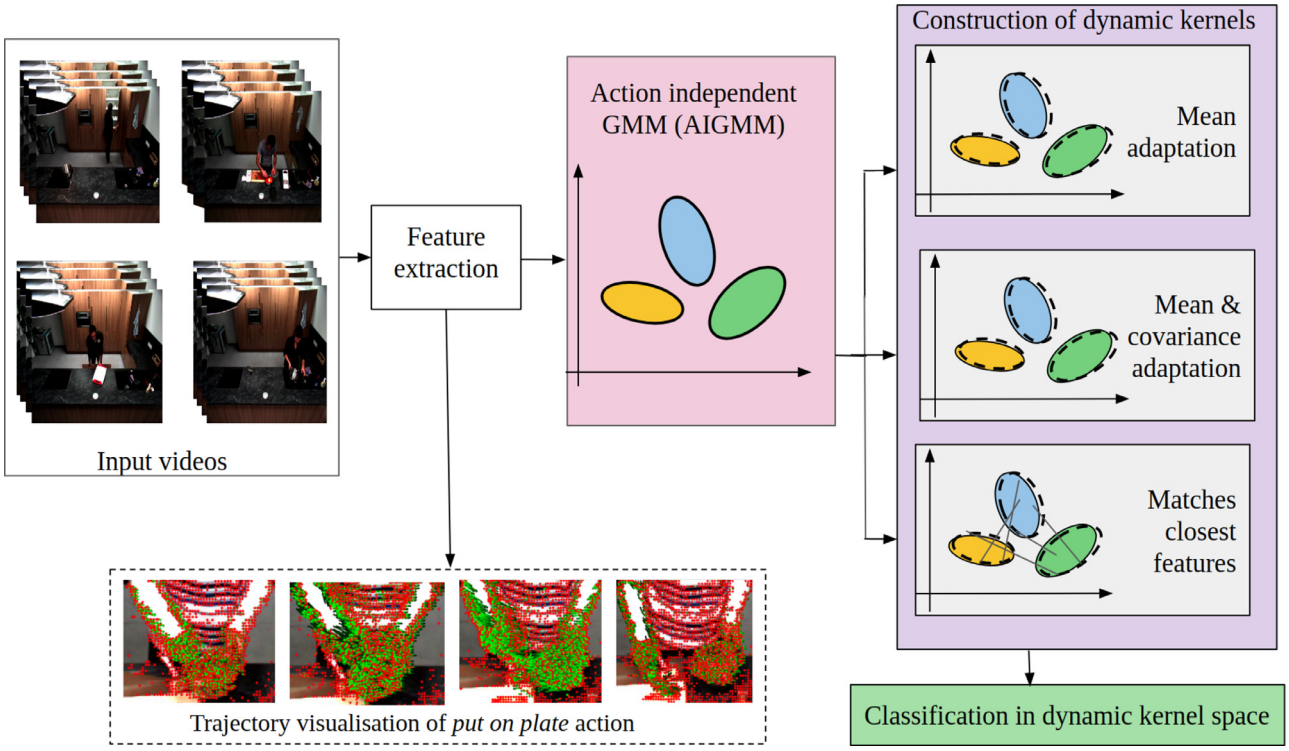


Fig. 3. Block diagram of the proposed approach for fine-grained action recognition (best viewed in colour).

3. Proposed approach

To overcome the above challenges, we propose a framework to obtain the efficient representation of fine-grained actions using dynamic kernels. The block diagram of the proposed approach is presented in the Fig. 3. Initially, spatio-temporal features, namely, histogram of optical flow (HOF) & motion boundary histogram (MBH) are extracted around the sampled feature points to capture the appearance and motion information, respectively. A large GMM, known as action independent GMM (AIGMM) is built on the features extracted from all the training video clips to model the correlations of subtle interactions among fine-grained actions. The statistics of learnt AIGMM are mapped onto dynamic kernel feature space to classify the fine-grained actions efficiently.

3.1. Feature extraction

Given an input video, the typical process of the recognition task is to extract spatio-temporal information for representing an action. We extract the spatio-temporal features around the densely sampled feature points to capture the crucial local motion information for modelling the subtle interactions in fine-grained actions. Feature points are densely sampled for various scales and are tracked across several consecutive frames to obtain the dense optical flow. This optical flow is considered to model the subtle interactions between the human and objects by capturing the absolute motion information among the sampled feature points. A trajectory is obtained by concatenating the points from consecutive frames. Descriptors, namely, HOG, HOF, & MBH features are extracted within the volume around the trajectory to capture the motion and appearance information. However, optical flow consists of background and camera motion that may bias the decision during action classification. Hence, we consider motion boundary histogram (MBH) features as it encodes the relative motion among pixels by computing spatial derivatives of optical flow leading to removal of the constant camera motion. Also, these features are robust to irregular motions and can capture the motion information efficiently [7]. An independent GMM is trained for each of the obtained descriptors separately to model the subtle variations among the fine-grained actions.

3.2. Action-independent Gaussian mixture model (AIGMM)

Conventionally, GMMs are used to encode the obtained spatio-temporal feature vectors by estimating the parameters of GMM using maximum likelihood estimation. Hence, we construct a single GMM using the training data of all actions known as Action-independent Gaussian mixture model (AIGMM) to model the attributes of different fine-grained actions. The AIGMM is represented as

$$p(\mathbf{x}_k|(w_q, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)) = \sum_{q=1}^{Q} w_q \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \tag{1}$$

where w_q are the mixture weights, satisfying the constraints, $0 \leq w_q \leq 1$, and $\sum_{q=1}^{Q} w_q = 1$. The $\boldsymbol{\mu}_q$ represents the mean and $\boldsymbol{\Sigma}_q$ denotes the covariance of the mixture q. The $\mathbf{x}_k$ denotes either a HOF or MBH descriptor. We train a separate AIGMM for each feature descriptor using Expectation maximization (EM) estimation. The EM algorithm estimates the parameters by maximizing the likelihood function given by Eq. (1). After training of AIGMM, each mixture of GMM is expected to capture an attribute of fine-grained actions. In order to increase the contribution of the attributes present in the video clip, the parameters of AIGMM are adpated using maximum aposterior (MAP) adaptation after observing each clip. The posterior probability of a AIGMM mixture, given the feature vector $\mathbf{x}_k$ is written as

$$p(q|\mathbf{x}_k) = \frac{w_q p(\mathbf{x}_k|q)}{\sum_{q=1}^{Q} w_q p(\mathbf{x}_k|q)}, \tag{2}$$

where w_q is the prior probability of the particular mixture q. The likelihood of the feature $\mathbf{x}_k$ coming from mixture q is represented

as $p(\mathbf{x}_k|q)$. The posterior probability $p(q|\mathbf{x}_k)$ and $\mathbf{x}_k$ are used to find the weight, mean, and covariance parameters [8] by

$$n_q(\mathbf{x}) = \sum_{k=1}^{K} p(q|\mathbf{x}_k), \tag{3}$$

$$\mathbf{F}_q(\mathbf{x}) = \frac{1}{n_q(\mathbf{x})} \sum_{k=1}^{K} p(q|\mathbf{x}_k)\mathbf{x}_k, \tag{4}$$

and

$$\mathbf{S}_q(\mathbf{x}) = \frac{1}{n_q(\mathbf{x})} \sum_{k=1}^{K} p(q|\mathbf{x}_k)\mathbf{x}_k^2. \tag{5}$$

respectively. The adapted weights, means, and covariance of each mixture q based on the posterior probability is given by

$$\hat{w}_q = \alpha n_q(\mathbf{x})/K + (1-\alpha)w_q, \tag{6}$$

$$\hat{\boldsymbol{\mu}}_q(\mathbf{x}) = \alpha \mathbf{F}_q(\mathbf{x}) + (1-\alpha)\boldsymbol{\mu}_q, \tag{7}$$

and

$$\hat{\boldsymbol{\Sigma}}_q(\mathbf{x}) = \alpha \mathbf{S}_q(\mathbf{x}) + (1-\alpha)(\boldsymbol{\Sigma}_q + \boldsymbol{\mu}_q^2) - \hat{\boldsymbol{\mu}}_q^2. \tag{8}$$

Here, α is the adaptating coefficient that maintains the balance between the old and new estimates.

3.3. Dynamic kernels

Kernel based approaches are proven to give better generalization performance for classification. Kernel methods construct an optimal linear solution by non-linearly transforming the input feature space to higher dimensional feature space. The performance of kernel methods depends on the choice of kernel function. The kernel functions i.e., dynamic kernels handle the varying length feature vectors by either mapping to fixed length patterns (probability based kernels) or selecting the best feature vectors (matching based kernels) to represent an action. Although recent methods exploit neural networks for classification, SVM is dominant when the training samples for each class are few in number and can be trained efficiently [39]. For a multi-class classification problem, the SVM based on the one-against-the-rest approach is used to discriminate the video clips of that class from video clips of all other classes. The SVM is a supervised learning model, which minimizes the objective function

$$J = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i^T, \mathbf{x}_j) - \sum_{i=1}^{n}\alpha_i, \tag{9}$$

where α_i are lagrange's multipliers and n is the number of video clips. $K(\mathbf{x}_i^T, \mathbf{x}_j)$ is the dynamic kernel function to obtain similarity between two vectors. We construct various dynamic kernel functions in the following sub-sections to achieve better discrimination. During the testing process, the decision function for the test example $\mathbf{x}_t$ is given by

$$f(\mathbf{x}_t) = sign\left(\sum_{i=1}^{m}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_t) + b\right). \tag{10}$$

The sign value of $f(\mathbf{x}_t)$ is used to determine the class of $\mathbf{x}_t$.

3.3.1. GMM supervector kernel (GMMSVK)

To determine the similarity between two fine-grained actions, we use the adapted means $\hat{\boldsymbol{\mu}}_q(\mathbf{x})$ of each clip obtained from Eq. (7) of AIGMM for constructing the GMMSVK kernel. The GMM supervector $\Phi_{sv}(\mathbf{x})$ is constructed by concatenating the GMM vectors $\boldsymbol{\varphi}_q(\mathbf{x}) = \left[\sqrt{w_q}\boldsymbol{\Sigma}_q^{-\frac{1}{2}}\hat{\boldsymbol{\mu}}_q(\mathbf{x})\right]^T$ of every mixture of AIGMM as

$$\Phi_{sv}(\mathbf{x}) = [\boldsymbol{\varphi}_1(\mathbf{x})^T, \boldsymbol{\varphi}_2(\mathbf{x})^T, \boldsymbol{\varphi}_3(\mathbf{x})^T, \ldots . \boldsymbol{\varphi}_Q(\mathbf{x})^T]^T. \tag{11}$$

Since, the dimension of GMM vector is of K for each of Q AIGMM mixtures, the dimension of GMM supervector leads to $Q \times K$. Finally, a GMMSVK for $\mathbf{x}_m$ & $\mathbf{x}_n$ examples is given by

$$K_{sv}(\mathbf{x}_m, \mathbf{x}_n) = \Phi_{sv}(\mathbf{x}_m)^T \Phi_{sv}(\mathbf{x}_n). \tag{12}$$

The calculation of GMMSVK involves computations of (i) $Q \times (L_m + L_n)$ for mean adaptation, (ii) $Q \times (K_l^2 + 1)$ for supervector, and (iii) K_s^2 for GMMSVK. Here, Q denotes the number of mixtures in AIGMM, L_m & L_n are the number of feature vectors for examples $\mathbf{x}_m$ & $\mathbf{x}_n$, respectively. K_l represents dimension of feature vector and K_s is dimension of supervector. Therefore, the total computation complexity of GMMSVK is $O(QL + QK_l^2 + K_s^2)$.

3.3.2. GMM mean interval kernel (GMMMIK)

Besides the adapted means, second-order statistics (covariances) provide the additional information about the distribution of fine-grained actions. Hence, the MIK exploits the first & second-order statistics of GMM by constructing the mean and covariance statistical vectors to capture the underlying useful information. The adapted means and covariances obtained from Eqs. (7) and (8) are utilized to build the GMMMIK kernel.

The GMM mean vector $\boldsymbol{\varphi}_q(\mathbf{x})$ for a video clip $\mathbf{x}$ is given by

$$\boldsymbol{\varphi}_q(\mathbf{x}) = \left(\frac{\hat{\boldsymbol{\Sigma}}_q(\mathbf{x}) + \boldsymbol{\Sigma}_q}{2}\right)^{-\frac{1}{2}}(\hat{\boldsymbol{\mu}}_q(\mathbf{x}) - \boldsymbol{\mu}_q). \tag{13}$$

The first term of Eq. (13) represents the degree of consistency of covariance matrices and second term gives the measure of deviation of means from the adapted means of AIGMM. The GMM mean vectors are concatenated to form a GMM mean supervector of dimension $Q \times K$ as

$$\Phi_{mv}(\mathbf{x}) = [\boldsymbol{\varphi}_1(\mathbf{x})^T, \boldsymbol{\varphi}_2(\mathbf{x})^T, \boldsymbol{\varphi}_3(\mathbf{x})^T, \ldots \boldsymbol{\varphi}_Q(\mathbf{x})^T]^T. \tag{14}$$

Finally, GMMMIK for examples $\mathbf{x}_m$ & $\mathbf{x}_n$ is given by

$$K_{mv}(\mathbf{x}_m, \mathbf{x}_n) = \Phi_{mv}(\mathbf{x}_m)^T \Phi_{mv}(\mathbf{x}_n). \tag{15}$$

The computation of GMMMIK involves computing mean adaptation, covariance adaptation, supervector computation, and GMMMIK computation. Hence, the total computation complexity of GMMMIK is given by $O(QL + Q(K_l^2 + K_l) + K_s^2)$. Due to the estimation of first & second-order statistics, the computational complexity of MIK is high.

3.3.3. Intermediate matching kernel (IMK)

Intermediate matching kernel (IMK) is one of the matching based dynamic kernels that computes the similarity between two fine-grained actions by finding the closest feature vectors. An IMK uses the set of virtual feature vectors denoted by $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots . \mathbf{v}_Q\}$ to match the set of local feature vectors. Now, IMK is constructed as

$$K_{imk}(\mathbf{x}_m, \mathbf{x}_n) = \Sigma_{q=1}^{Q} k(\mathbf{x}_{mq}^*, \mathbf{x}_{nq}^*). \tag{16}$$

Here, the feature vectors from examples $\mathbf{x}_m$ and $\mathbf{x}_n$ closest to q^{th} virtual feature vector $\mathbf{v}_q$ is given by

$$\mathbf{x}_{mq}^* = \text{argmin}_{x \in \mathbf{x}_m} D(\mathbf{x}, \mathbf{v}_q), \tag{17}$$

and

$$\mathbf{x}_{nq}^* = \text{argmin}_{x \in \mathbf{x}_n} D(\mathbf{x}, \mathbf{v}_q), \tag{18}$$

where $D(.,.)$ is the distance from virtual feature vector in $\mathbf{V}$ to a local feature vector in $\mathbf{x}_m$ or $\mathbf{x}_n$. The virtual feature vectors are

represented effectively by incorporating the mixtures of AIGMM $p(q|\mathbf{x}_k)$ as additional information using Eq. (2). The use of local feature vectors with mixtures of AIGMM induces information about mean vectors, covariance matrices, and mixture coefficients. The local feature vectors selected using the posterior probability for mixture q are given by

$$\mathbf{x}^*_{mq} = \text{argmax}_{x\in\mathbf{x}_m} p(q|\mathbf{x}_k), \tag{19}$$

and

$$\mathbf{x}^*_{nq} = \text{argmax}_{x\in\mathbf{x}_n} p(q|\mathbf{x}_k). \tag{20}$$

The computation of IMK involves measuring of $Q(L_m + L_n)$ distances resulting in total computational complexity of $O(QL)$. The total computations are less than the probabilistic based dynamic kernels when the number of mixtures Q is smaller than L_m & L_n.

4. Experiments

In the proposed approach, we calculate motion information by extracting descriptors such as HOF & MBH within the spatio-temporal volume of size $(2 \times 2 \times 3)$. The orientations of the descriptors are quantized into 9 bins resulting in a dimension of 108 $(2 \times 2 \times 3 \times 9)$ for HOF descriptor. In order to eliminate the background noise information captured during the extraction of optical flow, we compute spatial derivatives of the optical flow along the x and y directions. The orientation of these spatial derivatives are quantized into a 8 bin histogram for MBHx, MBHy separately resulting in the descriptor size of 96 $(2 \times 2 \times 3 \times 8)$ each. A single GMM is trained on the HOF and MBH descriptors separately for various mixtures ranging from 32, 64, 128, 256, and 512. The parameters of every GMM mixture are adapted using its posterior probability given the video clip. We consider top 20 posteriors to compute the adapted means and covariance matrices using Eqs. (7) and (8) for constructing the dynamic kernels because the posterior probabilities of AIGMM mixtures are mostly zero beyond the first 20 mixtures.

4.1. Datasets

The performance of the proposed approach is evaluated on 4 varieties of challenging fine-grained action datasets that are chosen from 3 different applications, namely, '*shopping*', '*medical surgeries*', and '*cooking*'. The shopping dataset consists of individual(s) performing shopping activity in grocery stores. The medical surgery dataset contains videos of a robotic arm performing surgeries. Finally, the cooking dataset comprises of individual(s) doing cooking activities. The videos of shopping and cooking datasets are collected from a single view-point with different illumination conditions. In contrast, medical surgery dataset is collected from two different cameras with subtle view-point variations. These datasets are challenging because of high intra-class variance, significant variability in the execution of tasks, and the subtle differences among fine-grained actions. The four challenging datasets are explained in detail in the following subsections.

4.1.1. Mitsubishi electric research laboratories (MERL) shopping

MERL dataset consists of shopping videos, recorded by a surveillance camera placed overhead [13]. The dataset contains videos of 32 different subjects performing shopping activity from grocery shelves. Each subject performs 5 different fine-grained actions, namely, '*reach from shelf (RFS)*', '*reach to shelf (RTS)* ', '*hand in shelf (HS)*', '*inspect product (IP)*', and '*inspect shelf (IS)*'. The dataset contains 79 training videos and 27 testing videos.

Table 1
Classification accuracy (%) of different dynamic kernels on various number of mixtures for MERL dataset.

	GMMSVK		GMMMIK		IMK	
	HOF	MBH	HOF	MBH	HOF	MBH
32	77.9	79.7	89.2	93.6	54.4	59.6
64	81.0	83.6	90.9	94.0	65.2	77.9
128	84.3	85.5	94.2	**96.5**	91.3	92.1
256	80.7	82.4	93.4	96.3	78.1	84.1
512	79.1	81.0	92.1	94.2	67.9	65.6

4.2. JHU-ISI gesture and skill assessment working set (JIGSAWS)

The dataset contains the kinematic and video data of robotic arm surgeries performed by 8 surgeons of varying surgical experience [40]. Each surgeon repeats 3 surgical tasks, namely, '*knot tying (KT)*', '*suturing (SUT)*', and '*needle passing (NP)*' five times resulting in variation in performing the same task. The videos are recorded from endoscopic cameras placed at the left and right side of the robotic arm to handle view-point variations. The dataset consists of 78, 56, and 72 videos of suturing, needle passing, and knot tying fine-grained actions, respectively. The dataset is split into training, testing based on the leave one user out (LOUO) setting, where the data of one subject out of 8 are considered for the test set and data of the remaining subjects for training [40].

4.3. Kitchen scene context-based gesture recognition (KSCGR)

Atsushi et al. [41] proposed the 'Actions for cooking eggs' (ACE) dataset and demonstrated its results in KSCGR. The dataset contains videos of 5 subjects cooking a meal with eggs in the kitchen. Each subject performs 8 fine-grained actions, namely, '*break*', '*mix*', '*bake*', '*turn*', '*cut*','*boil*', '*season*', and '*peel*'. It has 25 videos for training and 10 for testing each of 5 to 10 min duration. The challenges such as occlusion, large variation in performing the same task, low inter-class variance, etc., makes the dataset complex in recognising fine-grained actions.

4.4. Max Planck institute for informatics (MPII cooking2)

The dataset consists of an individual performing cooking activities in a constrained environment. Rohrbach et al. [4] introduced MPII cooking2 dataset to address the issues like low inter-class variability (e.g.: mix vs. stir), high intra-class variance (e.g.: cut tomato vs. cut pineapple), occlusion of objects while performing the action (e.g.: wash objects), low illumination, and presence of diverse objects (e.g.: knife, spiceholder, cutting board etc.). It contains 273 videos recorded by 30 individuals while performing 62 fine-grained actions, namely, '*cut dice*', '*cut stripes*', '*cut apart*', '*take lid*', '*put lid*', etc. The dataset is split into train and test based on subjects. Train set contains videos of 20 subjects, while the remaining 10 subjects are for testing.

4.5. Analysis of different dynamic kernels

The classification performance of various dynamic kernels, namely, GMM mean interval kernel (GMMMIK), GMM supervector kernel (GMMSVK), and intermediate matching kernel (IMK) are presented in Tables 1–4 on MERL, JIGSAWS, KSCGR, and MPII cooking2 datasets. The MIK-SVM built on 128 mixture GMM achieves the best performance for MERL, JIGSAWS, and KSCGR datasets due to the incorporation of first order and second order statistics of the AIGMM. Similarly, for MPII cooking2 dataset, 256 mixture GMM performs better than 128 mixtures. This is due to the fact that the AIGMM requires more mixtures to model all 62 fine-grained

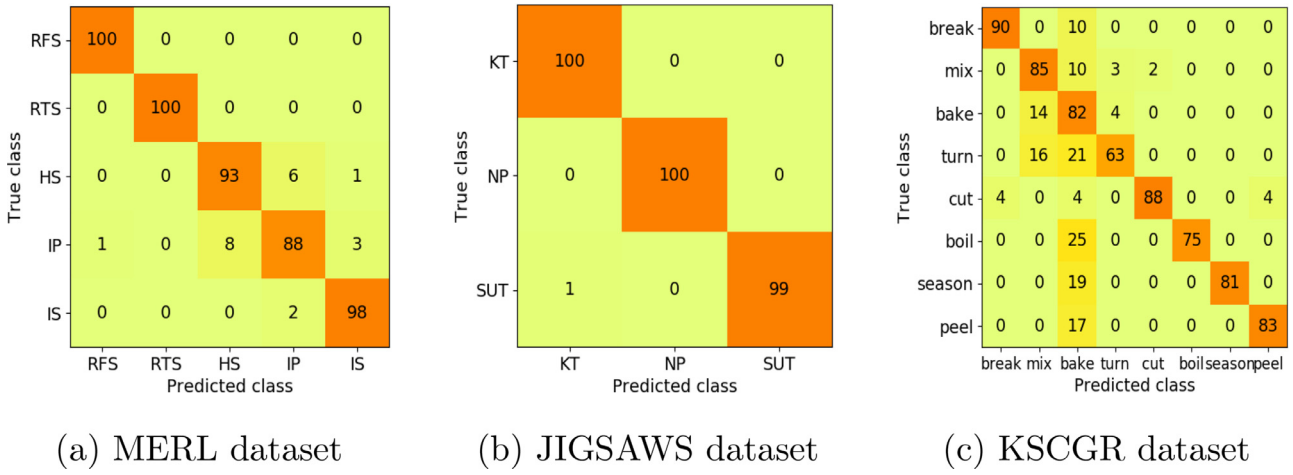


Fig. 4. Confusion matrices of best performance model (kernel - GMMMIK, 128 mixture GMM) for MERL, JIGSAWS, and KSCGR datasets.

Table 2
Classification accuracy (%) of different dynamic kernels on various number of mixtures for JIGSAWS dataset.

	GMMSVK		GMMMIK		IMK	
	HOF	MBH	HOF	MBH	HOF	MBH
32	93.8	94.5	94.7	94.7	92	94.3
64	94.2	95.6	95.1	99.5	93.3	95.7
128	95.4	96.1	98.5	**99.6**	94.2	96.0
256	92.3	94.1	97.5	97.6	93.1	96.4
512	91.0	93.4	95.3	96.4	92.4	94.5

Table 3
Classification accuracy (%) of different dynamic kernels on various number of mixtures for KSCGR dataset.

	GMMSVK		GMMMIK		IMK	
	HOF	MBH	HOF	MBH	HOF	MBH
32	26.6	31.4	66.7	79.8	40.5	60.0
64	33.6	33.4	66.7	80.0	63.9	67.9
128	42.7	44.6	75.4	**82.5**	80.2	81.3
256	39.4	40.4	75.4	79.4	79.3	71.0
512	35.1	36.9	74.2	77.4	74.6	62.7

Table 4
Classification accuracy (%) of different dynamic kernels on various number of mixtures for MPII cooking2 dataset.

	GMMSVK		GMMMIK		IMK	
	HOF	MBH	HOF	MBH	HOF	MBH
32	31.4	36.9	41.7	60.2	28.0	40.8
64	26.6	39.4	42.5	68.1	29.3	50.0
128	33.6	40.4	45.5	72.0	36.9	51.2
256	35.1	44.6	54.0	**75.6**	38.5	58.1
512	33.4	42.7	50.8	62.0	29.1	57.0

actions. Also, as the number of mixtures increased beyond 256, the classification performance is reduced. This may be due to the lack of adequate local information to capture the attributes of fine-grained actions. The small correlations of subtle interactions between the human and objects can be determined accurately because of large number of mixtures of AIGMM. This is essential in solving the high intra-class variability. Additionally, an AIGMM trained on the MBH descriptors performs better than the HOF due to its ability to capture the temporal information efficiently even in presence of irregular motions. The confusion matrices of the best performance model are presented in Figs. 4 and 5 for the MERL, JIGSAWS, KSCGR, and MPII cooking2 datasets, respectively. We also observe that the classification performance of each class is close to the overall classification performance, as the proposed model is able to capture the subtle interactions across the classes equally well. The scalability of the proposed approach is dependent on multiple factors like the number of AIGMM components, the number of local feature vectors for two clips to be compared, the dimension of the local feature vectors, the dimension of the supervector, and the number of training samples in a dataset.

Table 5
Performance comparison of proposed method with the state-of-the-art methods on MERL dataset.

Method	Accuracy (%)
MSB-RNN [13]	76.3
C3D [10]	90.8
Proposed GMMSVK	85.5
Proposed IMK	92.1
Proposed GMMMIK	**96.5**

Table 6
Performance comparison of proposed method with the state-of-the-art methods on JIGSAWS dataset.

Method	Accuracy (%)
Vector space model [49]	82.36
convnet [46]	93.06
CNN [45]	97.30
3D Conv Net [48]	98.30
Proposed GMMSVK	96.10
Proposed IMK	96.00
Proposed GMMMIK	**99.60**

4.6. Comparison with state-of-the-art methods

Tables 5–8 compare the performance of the proposed approach with state-of-the-art methods on MERL, JIGSAWS, KSCGR, and MPII cooking2 datasets, respectively. Existing approaches, namely, SIFT [43], trajectories [4], and IDT [4,44] encoded with Fisher vectors to model the fine-grained actions. Rohrbach et al [4] investigated the pose based approach to estimate the pose and track the body joints through the multiple frames. But this approach gives lower performance than the low-level features as it is based on the trajectories extracted from the joints, which are noisy. Fawaz et al. [45] and

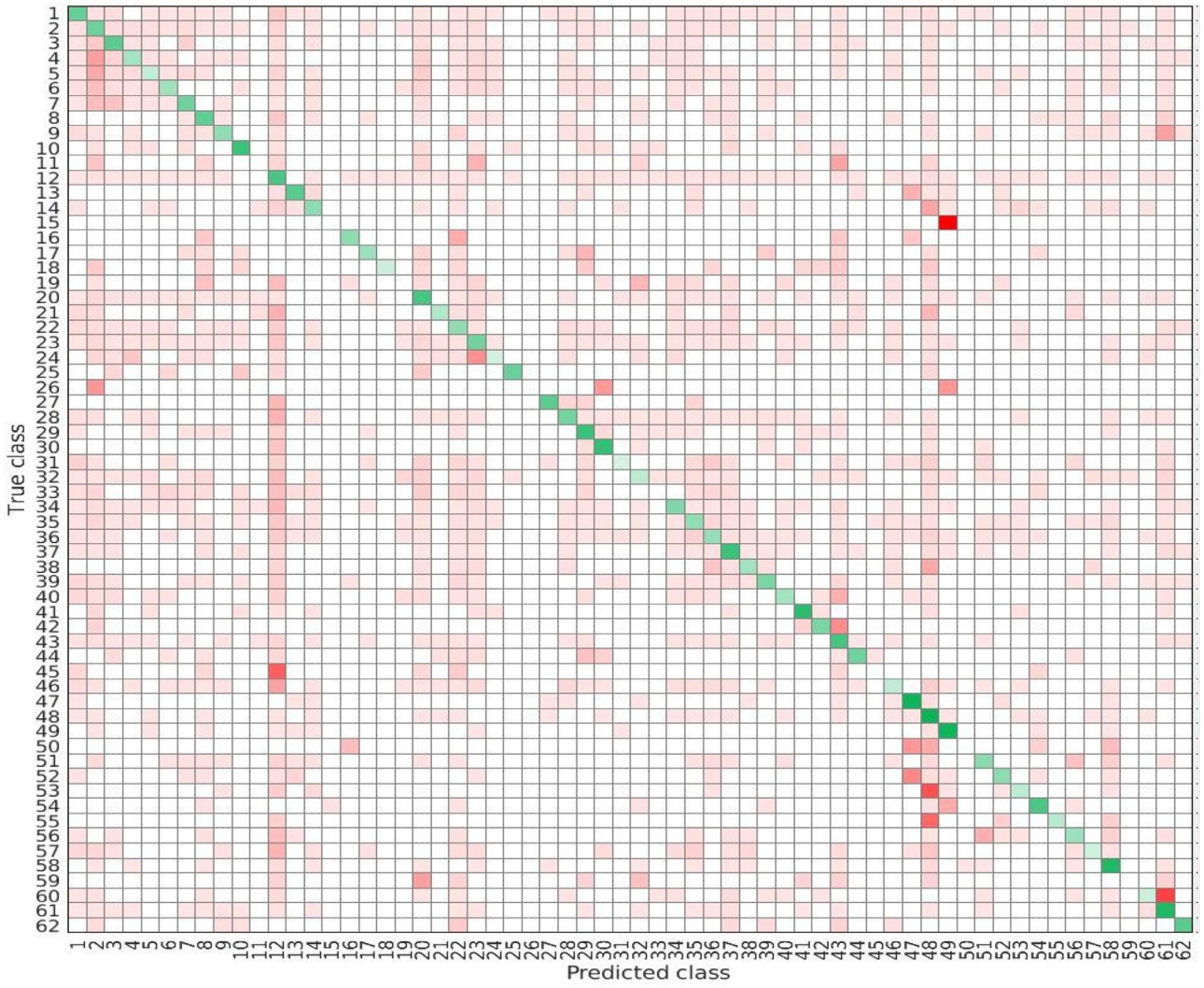


Fig. 5. Confusion matrix of best performance model (kernel - GMMMIK, 256 mixture GMM) for MPII cooking2 dataset. The indices of the classes are considered as in Rohrbach et al. [42].

Table 7
Performance comparison of proposed method with the state-of-the-art methods on KSCGR dataset.

Method	*F*-score
IDT-IFV-SVM [44]	0.76
RGB + OF + CNN + SVM [47]	0.70
RGB + OF + CNN + NN [47]	0.72
Proposed GMMSVK	0.41
Proposed IMK	0.82
Proposed GMMMIK	**0.85**

Table 8
Performance (%) comparison of proposed method with the state-of-the-art methods on MPII cooking2 dataset.

Method	mAP (%)
Pose-based approach [4]	24.1
Hand-cSIFT + Hand-Trajectories [4]	43.5
Dense trajectories [4]	44.5
Region-sequence CNN [30]	70.3
Proposed GMMSVK	45.0
Proposed IMK	58.0
Proposed GMMMIK	**75.0**

Wang et al. [46] exploited the CNNs to extract the discriminative patterns of fine-grained actions. However, the conventional CNNs capture only spatial information while ignoring the prominent motion cues. To overcome this, multi-stream networks [13,30,47] are employed to capture the spatio-temporal information of fine-grained actions efficiently. Yet, these networks are trained independently on RGB frames and optical flow to learn the appearance and motion information, respectively, ignoring the need for interactions among the multiple streams. In order to capture the spatial and temporal cues laterally, 3D-CNNs [10,48] are investigated to classify the video snippets extracted from untrimmed videos. However, the above mentioned approaches are computationally complex, require large amount of training data, and do not generalize on the smaller datasets. It can be observed from the Tables 5–8 that the proposed approach exhibits better performance than the existing methods by capturing the local and global context of motion dynamics in fine-grained actions efficiently. We compare our approach with the existing methods using accuracy, mAP and F-score evaluation metrics.

5. Conclusion

In this paper, we propose a novel approach for an efficient and compact representation of fine-grained actions which are of varying length by exploring various dynamic kernels. An action independent GMM (AIGMM) is trained on the extracted spatio-temporal features to capture the local similarities among the fine-grained actions. A dynamic kernel representation that incorporated the first & second-order statistics of the trained GMM is shown to recognise the fine-grained actions efficiently. This is because of the fact that these statistics effectively model the global information by handling the actor, view-point, and illumination variations. We demonstrate the generalization of the proposed approach by evaluating on 4 wide varieties of datasets, namely, MERL, JIGSAWS, KSCGR, and MPII cooking2. The proposed approach shows that the dynamic kernels are suitable choice for fine-grained action recognition.

Although intermediate matching kernel (IMK) demonstrates low computational complexity, its classification performance is limited when compared to probabilistic based dynamic kernels. It is due to the fact that the selected local feature vectors based on the GMM mixtures are common for all classes. In future, this work can be extended by constructing the feature vectors that are specific to respective classes for efficient representation of fine-grained actions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would express their thanks to Dr. D Rajesh reddy, Scientist-SE, ADRIN, Dept. of Space, Secunderabad for insightful discussions.

References

[1] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, A.G. Hauptmann, DevNet: a deep event network for multimedia event detection and evidence recounting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2568–2577.
[2] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, J. Huang, End-to-end learning of motion representation for video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6016–6025.
[3] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, J. Huang, Weakly supervised dense event captioning in videos, (2018), arXiv preprint arXiv:1812.03849.
[4] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, B. Schiele, Recognizing fine-grained and composite activities using hand-centric features and script data, Int. J. Comput. Vis. (IJCV) 119 (3) (2016) 346–373.
[5] I. Laptev, On space-time interest points, Int. J. Comput. Vis. (IJCV) 64 (2) (2005) 107–123.
[6] S. Paul, A. Saad, S. Mubarak, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM International Conference on Multimedia, 2007, pp. 357–360.
[7] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011, IEEE, 2011, pp. 3169–3176.
[8] D. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, in: Digital Signal Process., 10, 2000, pp. 19–41.
[9] T. Zhigang, X. Wei, Q. Qianqing, P. Ronald, V.R. C, L. Baoxin, Y. Junsong, Multi-stream CNN: learning representations based on human-related regions for action recognition, Pattern Recognit. 79 (2018) 32–43.
[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[11] B. Ni, X. Yang, S. Gao, Progressively parsing interactional objects for fine grained action detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1020–1028.
[12] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML), 28, 2013, pp. 1310–1318.
[13] B. Singh, T.K. Marks, M. Jones, O. Tuzel, M. Shao, A multi-stream bi-directional recurrent neural network for fine-grained action detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1961–1970.
[14] V. Thenkanidiyoor, D. AD, C. Chandra Sekhar, Dynamic kernels based approaches to analysis of varying length patterns in speech and image processing tasks, in: Pattern Recognition and Big Data, World Scientific, 2017, pp. 407–485.
[15] I. Alexandros, T. Anastastios, P. Ioannis, Discriminant bag of words based representation for human action recognition, Pattern Recognit. Lett. 49 (2014) 185–192.
[16] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1704–1716.
[17] S. Manel, M. Mahmoud, A.C. Ben, Human action recognition based on multi-layer fisher vector encoding method, Pattern Recognit. Lett. 65 (2015) 37–43.
[18] Y. Li, W. Li, V. Mahadevan, N. Vasconcelos, Vlad3: encoding dynamics of deep features for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1951–1960.
[19] W. Hao, Z. Zhang, Spatiotemporal distilled dense-connectivity network for video action recognition, Pattern Recognit. 92 (2019) 13–24.
[20] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[21] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, (2014), arXiv preprint arXiv:1406.2199.
[22] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.
[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
[24] Y. Hao, Y. Chunfeng, L. Bing, D. Yang, X. Junliang, H. Weiming, M.S. J, Asymmetric 3D convolutional neural networks for action recognition, Pattern Recognit. 85 (2019) 1–12.
[25] J. Li, X. Liu, M. Zhang, D. Wang, Spatio-temporal deformable 3D convnets with attention for action recognition, Pattern Recognit. 98 (2020) 107037.
[26] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7083–7093.
[27] Y. Zhou, B. Ni, R. Hong, M. Wang, Q. Tian, Interaction part mining: a mid-level approach for fine-grained action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3323–3331.
[28] M. Cheng, Z. Zhang, W. Lin, P. Torr, Bing: binarized normed gradients for objectness estimation at 300 fps, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3286–3293.
[29] F. Liu, L. Zhao, X. Cheng, Q. Dai, X. Shi, J. Qiao, Fine-grained action recognition by motion saliency and mid-level patches, Appl. Sci. 10 (8) (2020) 2811.
[30] M. Ma, N. Marturi, Y. Li, A. Leonardis, R. Stolkin, Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos, Pattern Recognit. 76 (2018) 506–521.
[31] Y. Zhu, G. Liu, Fine-grained action recognition using multi-view attentions, Vis. Comput. 36 (9) (2020) 1771–1781.
[32] T. Han, H. Yao, W. Xie, X. Sun, S. Zhao, J. Yu, TVENet: temporal variance embedding network for fine-grained action representation, Pattern Recognit. 103 (2020) 107267.
[33] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[34] A.D. Dileep, C.C. Sekhar, GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines, IEEE Trans. Neural Netw. Learn. Syst. 25 (8) (2013) 1421–1432.
[35] A.D. Dileep, C.C. Sekhar, Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines, Speech Commun. 57 (2014) 126–143.
[36] S. Boughorbel, J.P. Tarel, N. Boujemaa, The intermediate matching kernel for image local features, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., 2, IEEE, 2005, pp. 889–894.
[37] W.M. Campbell, D.E. Sturim, D.A. Reynolds, Support vector machines using GMM supervectors for speaker verification, IEEE Signal Process. Lett. 13 (5) (2006) 308–311.
[38] C.H. You, K.A. Lee, H. Li, A GMM supervector kernel with the Bhattacharyya distance for SVM based speaker recognition, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 4221–4224.
[39] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.
[40] Y. Gao, S.S. Vedula, C.E. Reiley, N. Ahmidi, B. Varadarajan, H.C. Lin, L. Tao, L. Zappella, B. Béjar, D.D. Yuh, et al., JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling, in: MICCAI workshop: M2cai, 3, 2014, p. 3.
[41] A. Shimada, K. Kondo, D. Deguchi, G. Morin, H. Stern, Kitchen scene context based gesture recognition: acontest in ICPR2012, in: International Workshop on Depth Image Analysis and Applications, Springer, 2012, pp. 168–185.
[42] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1194–1201.

[43] X. Sun, M. Chen, A. Hauptmann, Action recognition via local descriptors and holistic features, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 58–65.
[44] B. Ni, V.R. Paramathayalan, P. Moulin, Multiple granularity analysis for fine–grained action detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 756–763.
[45] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, Evaluating surgical skills from kinematic data using convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 214–221.
[46] Z. Wang, A.M. Fey, Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery, Int. J. Comput. Assist. Radiol. Surg. 13 (12) (2018) 1959–1970.
[47] R.L. Granada, J. Monteiro, R.C. Barros, F.R. Meneguzzi, A deep neural architecture for kitchen activity recognition, in: Proceedings of the 30th Florida Artificial Intelligence Research Society Conference, 2017, Estados Unidos., 2017.
[48] I. Funke, S.T. Mees, J. Weitz, S. Speidel, Video-based surgical skill assessment using 3D convolutional neural networks, Int. J. Comput. Assist. Radiol. Surg. 14 (7) (2019) 1217–1225.
[49] G. Forestier, F. Petitjean, P. Senin, F. Despinoy, P. Jannin, Discovering discriminative and interpretable patterns for surgical motion analysis, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2017, pp. 136–145.

Yenduri Sravani received the B.Tech degree from R.V.R & J.C. College of Engineering, Chowdavaram, Guntur, India, in 2015, and M.Tech degree from University College of Engineering, JNTU Kakinada, India, in 2017. She is currently working towards the Ph.D. degree in the department of computer science, IIT Hyderabad. Her current research interests include video content analysis, action recognition, fine-grained action recognition, and computer vision.

Nazil Perveen received the B.E. degree in computer science and engineering from Guru Ghasidas University, India, in 2009, the M.Tech. degree in computer technology from the National Institute of Technology Raipur, India, in 2012, and the Ph.D. degree in applied machine learning from Computer Science and Engineering, Indian Institute of Technology Hyderabad, India, in May 2020. Her research interests include pattern recognition, deep learning, human behaviour analysis, emotion recognition, and medical image analysis.

Chalavadi Vishnu received a B.Tech degree in Computer Science Engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2016. Received M.Tech degree from Indian Institute of Technology Hyderabad (IITH), India in CSE, in 2018. Currently pursuing a Ph.D. at IIT Hyderabad in the Department of Computer Science Engineering. Research interests include learning graph representations on video activities, deep learning for drones, and autonomous vehicles.

Dr. Krishna Mohan C (Member, IEEE) received the B.Sc.Ed. degree from the Regional Institute of Education, India, in 1988, the M.C.A. degree from the S. J. College of Engineering, India, in 1991, the M.Tech. degree in system analysis and computer applications from the National Institute of Technology Surathkal, India, in 2000, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, India, in 2007. He is currently a Professor with the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India. His research interests include video content analysis, pattern recognition, and neural networks.