Fine-grained Action Recognition using Attribute Vectors

Sravani Yenduri, Nazil Perveen, Vishnu Chalavadi and C. Krishna Mohan
Indian Institute of Technology Hyderabad, Kandi, Sangareddy, Telangana, 502285, India

Keywords: Spatio-temporal Features, Gaussian Mixture Model (GMM), Maximum A Posterior (MAP) Adaptation, Factor Analysis, Fine-grained Action Recognition.

Abstract: Modelling the subtle interactions between human and objects is crucial in fine-grained action recognition. However, the existing methodologies that employ deep networks for modelling the interactions are highly supervised, computationally expensive, and need a vast amount of annotated data for training. In this paper, a framework for an efficient representation of fine-grained actions is proposed. First, spatio-temporal features, namely, histogram of optical flow (HOF), and motion boundary histogram (MBH) are extracted for each input video as these features are more robust to irregular motions and capture the motion information in videos efficiently. Then a large Gaussian mixture model (GMM) is trained using the maximum a posterior (MAP) adaption, to capture the attributes of fine-grained actions. The adapted means of all mixtures are concatenated to form an attribute vector for each fine-grained action video. This attribute vector is of large dimension and contains redundant attributes that may not contribute to the particular fine-grained action. So, factor analysis is used to decompose the high-dimensional attribute vector to a low-dimension in order to retain only the attributes which are responsible for that fine-grained action. The efficacy of the proposed approach is demonstrated on three fine-grained action datasets, namely, JIGSAWS, KSCGR, and MPII cooking2.

1 INTRODUCTION

The fundamental task of action recognition is to distinguish various human actions performed in a given video. Human action recognition has gained interest in recent years because of its potential in applications like surveillance videos, video retrieval, human-robot interaction, and autonomous driving vehicles. Despite the intensive progress in action recognition, the existing state-of-the-art methods recognize only full body activities like jumping, waving, etc. But these methods are unable to differentiate between actions such as cut & peel, take out from cupboard & take out from fridge, cut dice & cut stripes etc. These actions are visibly similar to each other and have high inter-class similarity as shown in Figure 1, which are called fine-grained actions. For instance, recognizing similar actions in a cooking activity like cut, cut dice, cut apart, peeling constitute fine-grained action recognition. Recognizing human actions in finer detail has increased research interest because of its applications in human-computer interaction, video description, and surveillance videos. Fine-grained action recognition is very challenging when compared to action recognition due to high inter-class similarity, low intra-class similarity, presence of diverse objects, large variation in performing the same task, occlusion, and viewpoint variations.

(a) cut (b) peel

Figure 1: Fine-grained actions in cooking activity (Rohrbach et al., 2016).

Existing methods model the interaction between human and objects by detecting the objects explicitly. The explicit object detection methods require large annotated data, and cannot detect objects in low illumination conditions. To overcome this problem, Ni (Ni et al., 2016) employed LSTMs to recognise the object-specific actions by consolidating the object detections. Also, bi-directional long short term memory (Bi-LSTM) (Singh et al., 2016) is used to model the long-term temporal association between the human and objects without the need for explicit object

Yenduri, S., Perveen, N., Chalavadi, V. and Mohan, C.
Fine-grained Action Recognition using Attribute Vectors.
DOI: 10.5220/0010828700003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 134-143
ISBN: 978-989-758-555-5; ISSN: 2184-4321

detection. The aforementioned approaches are computationally complex while calculating deep features for given videos.

The holistic approach for better representation of fine-grained actions with low computational complexity is to encode dense trajectories with Fisher vector representation. However, the use of only low-level features for classification will restrict the local information which is important in recognising fine-grained actions. So there is a need for better discriminative representation using these low-level features to model the human-object interactions, as the low-level features are computationally less expensive.

In this paper, we propose an approach to obtain the universal representation for each fine-grained action without explicit object detection. The main aim of the proposed work is to capture the attributes that can model the interactions effectively in a single model. Here, attributes are the units that form a fine-grained action. For instance, *cut* action in cooking activity is described as a sequence of attributes such as right-wrist retraction, left-hand rotate etc. In the proposed approach, we train a large Gaussian mixture model (GMM) to model the attributes of all fine-grained actions. Then an attribute vector is formed by concatenating the means which are adapted using maximum a posteriori adaptation (MAP) for each fine-grained action. This attribute vector is of high-dimension and consists of the redundant attributes that do not contribute to a particular fine-grained action. So, this attribute vector is decomposed to low-dimensional one using factor analysis for efficient representation of the features. We evaluate our method on three fine-grained action datasets, namely, JHU-ISI gesture skill & assessment working set (JIGSAWS), kitchen scene context-based gesture recognition (KSCGR), and Max Planck Institute for Informatics (MPII) cooking2 datasets. The main contributions of the proposed method are:

- We propose a framework to represent fine-grained actions using attribute modelling where annotation of objects is not needed explicitly.
- We obtain a low-dimensional feature representation of each video clip for better discrimination of fine-grained actions.
- Demonstration of proposed approach on 3 wide variety of datasets constituting fine-grained actions. These datasets include cooking activity and robotic arm surgeries as fine-grained actions to be recognised.

2 RELATED WORK

In order to overcome the existing challenges, extensive research has been carried out to recognise actions in trimmed videos. In the past decade, research in action recognition has evolved from traditional hand crafted methods to current deep learning methods. Traditional action recognition methods explored several hand crafted features, namely, spatio-temporal interestpoints (STIP) (Ivan, 2005), improved dense trajectory (IDT) (Wang et al., 2011), etc to represent actions in a video. These features are usually extracted by tracking the interest points that are either densely sampled or detected by 3D harris corner detector, throughout the video. Descriptors, namely, HOG, HOF, and MBH are extracted around these interest points and encoded using different feature encoding techniques such as Fisher vector (Manel et al., 2015) , vector of locally aggregated descriptors (VLAD) (Herve et al., 2011), and Bag of words (Alexandros et al., 2014) to classify actions present in a video. Maria (Maria and Joan, 2018) proposed an approach to improve the performance of IDT by incorporating a new feature namely temporal templates. These temporal templates are constructed by computing three different projections for an input video. Feature descriptors of each projection are encoded using Fisher vectors. The Fisher vectors from these three projections are integrated by sum pooling and are fed to SVM to classify actions.

Motivated by the success of deep learning methods in various vision tasks such as image classification, object recognition, and segmentation, several CNN based approaches have been presented for action recognition. Andrej (Andrej et al., 2014) has investigated several fusion techniques to incorporate the temporal information, as conventional CNN captures only the spatial information from the RGB frames. Later two-stream networks (Heeseung et al., 2018; Yamin et al., 2018) are proposed, to learn the spatial and motion information by individual streams whose input is RGB frames and optical flow of few consecutive frames, respectively. Encouraged from a two-stream network, Zhigang (Zhigang et al., 2018), proposed a multi-stream CNN to classify actions. First, the region of interest is extracted from each frame by using a motion saliency measure. These regions of interest are considered to contain discriminative information that are essential to classify actions. The inputs to multi-stream are the images cropped to the region of interest and an entire RGB image to incorporate both local and global spatial information, respectively. However, the 2D-CNN, multi-stream networks are efficient in extracting only spatial informa-

tion, but tend to ignore the discriminative motion information necessary for classifying actions.

In order to overcome this limitation, 3D-CNN models (Tran et al., 2015) are introduced to capture spatio-temporal information of an action. Nonetheless, 3D models are difficult to train and computationally expensive. Later, Wang et al. (Wang et al., 2016) 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. Hao (Hao et al., 2019) proposed an asymmetric 3D convolutional network to reduce the number of parameters and computational complexity. Similarly, temporal shift module (TSM) (Lin et al., 2019) 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 object which are crucial in fine-grained actions.

Zhou (Zhou et al., 2015) proposed a mid-level approach to model the interactions between human and objects, by generating the discriminative interaction regions. This method does not need explicit object detection and thus reduces the human labor for annotation. The interaction regions are generated using the BING proposal tool (Cheng et al., 2014) and these are tracked based on the appearance, motion, and spatial overlap to form a graph. The resultant graph is divided into sub-graphs by graph segmentation algorithm to classify fine-grained actions, where these sub-graphs represent the human object interaction parts.

Singh (Singh et al., 2016) proposed a multi-stream bi-directional recurrent neural network (MSB-RNN) for localizing a fine-grained action temporally and spatially in each frame. Similarly, Miao (Miao et al., 2018), addressed the issues of both coarse and fine-grained action recognition by introducing a region based six stream CNN model. Firstly, prominent human poses and positions are detected in the video sequence and are cropped to different scale regions to obtain the richer spatio-temporal information. These cropped regions are fed to 6 independent CNNs as inputs and the obtained feature descriptors are concatenated to classify fine-grained actions. This framework efficiently recognises fine-grained actions by claiming that the spatial regions contain better discriminative information. But, it overlooks the fact that the temporal information is also significant in fine-grained action recognition. For example, *take out from cupboard* and *put in cupboard* fine-grained actions can be distinguished only by considering motion information.

However, the limitations of existing approaches are: (i) highly supervised, need a large amount of annotated data and (ii) computationally expensive when extracting the deep features of long duration videos.

3 PROPOSED APPROACH

The block diagram for the proposed approach is presented in Figure 2. First, the spatio-temporal features like the HOF and MBH are extracted for each input video. Then a large GMM is trained using the maximum aposteriori (MAP) adaptation, to capture the attributes of fine-grained actions. The adapted means of all mixtures of a large GMM are concatenated to form a high-dimensional attribute vector. Finally, the obtained feature vector is reduced to a low-dimensional attribute vector using factor analysis to retain only the attributes responsible for the fine-grained action. The above framework is described in detail in the following sub-sections.

3.1 Feature Extraction

Initially, feature points are densely sampled on a grid with a step size of 5 pixels on each spatial scale separately. Trajectories are extracted densely for 8 spatial scales and the main objective of the dense trajectories is to track the points throughout the video (Wang et al., 2011). The tracked position of the point $Q_t = (x_t, y_t)$ in the frame F_{t+1} is obtained by using a median filtering kernel K on dense optical flow field o_t

$$Q_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (K \times o_t)|_{x_t, y_t}. \quad (1)$$

Trajectories are formed by concatenating the points of subsequent frames $(Q_t, Q_{t+1}, Q_{t+2}, ...)$. Trajectories drift away from their initial position during the tracking process, so the trajectory length is confined to 15 frames. The descriptors are computed within a spatio-temporal volume of size $M \times M$ pixels and 15 frames long to find the motion information. The orientations of the descriptors are quantized into 8 bins. Further zero bin is added for HOF (9 bins). The HOG descriptor is of size 96 $(2 \times 2 \times 3 \times 8)$ and HOF is of 108 descriptor size $(2 \times 2 \times 3 \times 9)$. Optical flow consists of background camera motion and the motion of camera may bias the decision of action classification. In order to overcome this limitation, we consider motion boundary histogram (MBH) features as it computes spatial derivatives of optical flow leading to removal of the constant camera motion. The orientation of the spatial derivatives is quantized into

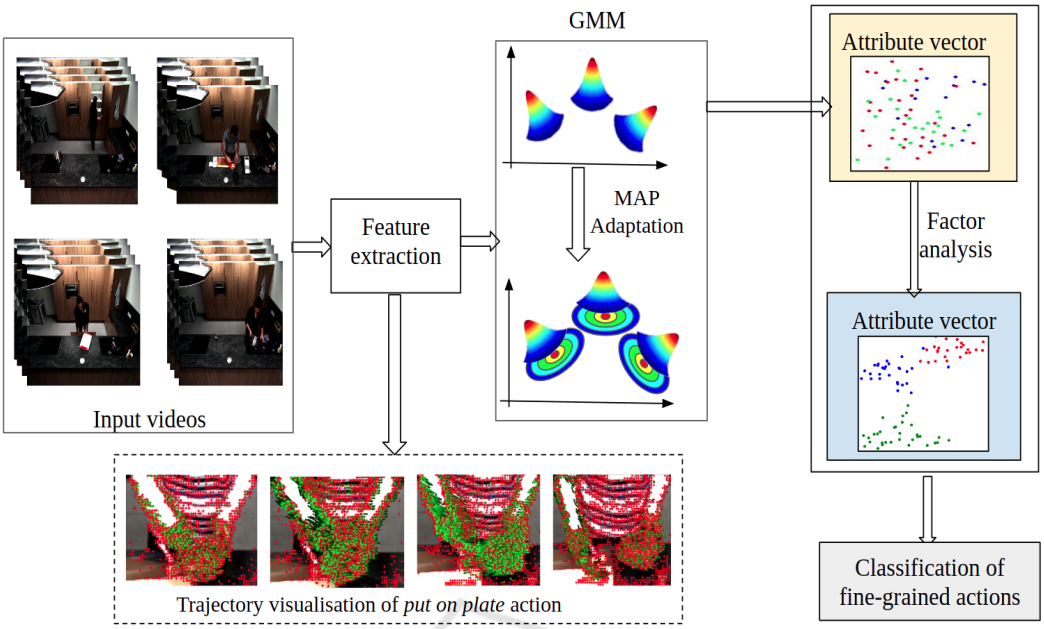


Figure 2: Systematic representation of the proposed approach (best viewed in color).

8 bin histogram for MBHx, MBHy separately leading to the descriptor size of 96 $(2 \times 2 \times 3 \times 8)$ each. Thus, the obtained feature descriptors are used to model the large GMM for each descriptor separately.

3.2 Gaussian Mixture Models (GMM)

A video clip is considered to be a random process whose distribution is assumed to be Gaussian. In order to find the similarity among fine-grained action clips, parameters of the Gaussian distribution have to be estimated. These parameters are estimated by training a GMM for each fine-grained action. Thus a single large GMM is trained because training a GMM for each fine-grained action is challenging when there is a vast number of actions. The GMM can be represented as

$$p(\mathbf{x}_k) = \sum_{q=1}^{Q} w_q \mathcal{N}(\mathbf{x}_k | \boldsymbol{\mu}_q, \boldsymbol{\sigma}_q), \tag{2}$$

where w_q are the mixture weights, which satisfy the constraints, $0 \leq w_q \leq 1$, and $\sum_{q=1}^{Q} w_q = 1$. The mean and covariance of the mixture q are given by $\boldsymbol{\mu}_q$ and $\boldsymbol{\sigma}_q$, respectively. Feature vector $\mathbf{x}_k$ belongs to $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$ of a video clip $\mathbf{x}$. The $\mathbf{x}_k$ can be a HOF or MBH descriptor and a separate large GMM is trained for each feature descriptor using Expectation maximisation (EM) estimation. After training the GMM, we assume that each component of GMM captures an attribute of fine-grained actions. MAP adaptation is used to obtain the probability distribution function *(pdf)* that describes the clip.

3.3 Attribute Vector Representation

The posterior probability of an attribute, given the feature vector $\mathbf{x}_k$ is written as

$$p(q|\mathbf{x}_k) = \frac{w_q p(\mathbf{x}_k|q)}{\sum_{q=1}^{Q} w_q p(\mathbf{x}_k|q)}, \tag{3}$$

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 likelihood $p(q|\mathbf{x}_k)$, and $\mathbf{x}_k$ are used to find the weight, and mean parameters (Reynolds et al., 2000) also known as zeroth and first order Baum-Welch statistics, given by

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

and

$$F_q(\mathbf{x}) = \frac{1}{n_q(\mathbf{x})} \sum_{k=1}^{K} p(q|\mathbf{x}_k)\mathbf{x}_k, \tag{5}$$

respectively. The adapted means and weights of each mixture q is given by

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

and

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

The obtained adapted means of each mixture q are concatenated to form $QK \times 1$ high dimensional attribute vector, i.e.,

$$\mathbf{A}(\mathbf{x}) = [\hat{\boldsymbol{\mu}_1}\hat{\boldsymbol{\mu}_2}...\hat{\boldsymbol{\mu}_Q}]^t. \quad (8)$$

This high-dimensional attribute vector consists of the attributes that do not contribute to the video clip, which results in close to zero Baum-Welch statistics. So, we use an appropriate decomposition method to obtain the efficient low-dimensional attribute vector. The high dimensional attribute vector is decomposed as $\mathbf{A} = \mathbf{m} + \mathbf{V}\mathbf{r}$, where $\mathbf{m}$ is a vector independent of viewpoint variation, $\mathbf{V}$ is known as variability matrix of size $QK \times l$ and $\mathbf{r}$ is an l-dimensional random vector having Gaussian distribution. This random vector is referred to as a low-dimensional attribute vector and is given by the posterior distribution $P(\mathbf{r}|\mathbf{x})$ i.e.,

$$p(\mathbf{r}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{r})\mathcal{N}(0,1),$$
$$and$$
$$p(\mathbf{r}|\mathbf{x}) \propto exp\Big(-\frac{1}{2}(\mathbf{r}-\mathbf{H}(\mathbf{x}))^t\mathbf{M}(\mathbf{x})(\mathbf{r}-\mathbf{H}(\mathbf{x}))\Big), \quad (9)$$

where $\mathbf{H}(\mathbf{x}) = \mathbf{M}^{-1}(\mathbf{x})\mathbf{V}^t\boldsymbol{\Sigma}^{-1}\hat{\mathbf{A}}(\mathbf{x})$, $\hat{\mathbf{A}}(\mathbf{x})$ is the centred vector, and $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $QK \times QK$. The mean of the adapted GMM is given by

$$\hat{\mathbf{F}}_q(\mathbf{x}) = \sum_{k=1}^{K} p(q|\mathbf{x}_k)(\mathbf{x}_k - \boldsymbol{\mu}_q). \quad (10)$$

The $\hat{\mathbf{A}}(\mathbf{x})$ is formed by concatenating the first-order statistics as $\hat{\mathbf{A}}(\mathbf{x}) = [\hat{\mathbf{F}}_1(\mathbf{x})\hat{\mathbf{F}}_2(\mathbf{x}).....\hat{\mathbf{F}}_Q(\mathbf{x})]^t$. The matrix $\mathbf{M}(\mathbf{x})$ is defined as $\mathbf{M}(\mathbf{x}) = \mathbf{I} + \mathbf{V}^t\boldsymbol{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{V}$, where $\mathbf{N}(\mathbf{x})$ is a diagonal matrix with $n_q(\mathbf{x})\mathbf{I}$ of $QK \times QK$ dimension and $\mathbf{I}$ is the identity matrix. The mean and covariance matrix from Equation 7 are given by $E[\mathbf{r}(\mathbf{x})] = \mathbf{M}^{-1}(\mathbf{x})\mathbf{V}^t\boldsymbol{\Sigma}^{-1}\hat{\mathbf{A}}(\mathbf{x})$, and $Cov(\mathbf{r}(\mathbf{x}),\mathbf{r}(\mathbf{x})) = \mathbf{M}^{-1}(\mathbf{x})$, respectively. EM algorithm is employed to estimate the mean and covariance iteratively in the E-step and to update $\mathbf{V}$, $\boldsymbol{\Sigma}$ in the M-step. In E-step, $\mathbf{m}$ and $\boldsymbol{\Sigma}$ are initialized with GMM mean and covariance, respectively. In M-step, $\mathbf{V}$ is obtained by solving $\sum_{\mathbf{x}}\mathbf{N}(\mathbf{x})\mathbf{V}E[\mathbf{r}(\mathbf{x})\mathbf{r}^t(\mathbf{x})] = \sum_{\mathbf{x}}\hat{\mathbf{A}}(\mathbf{x})E[\mathbf{r}^t(\mathbf{x})]$, which results in l linear equations. The residual matrix is given by

$$\boldsymbol{\Sigma}_q = \frac{1}{n_q\mathbf{x}}\Big(\sum_x \hat{\mathbf{S}}_q(\mathbf{x}) - \mathbf{M}_q\Big), \quad (11)$$

where $\mathbf{M}_q$ is the q^{th} diagonal block of the $QK \times QK$ matrix and $\hat{\mathbf{S}}_q(\mathbf{x})$ is the second-order statistics given by

$$\hat{\mathbf{S}}_q(\mathbf{x}) = diag\Big(\sum_{k=1}^{K} p(q|\mathbf{x}_k)(\mathbf{x}_k - \boldsymbol{\mu}_q)(\mathbf{x}_k - \boldsymbol{\mu}_q)^t\Big). \quad (12)$$

After the final estimation of $\mathbf{V}$ and Σ matrices, the attribute vector for a given clip is written as

$$\mathbf{r}(\mathbf{x}) = (I + \mathbf{V}^t\boldsymbol{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{V})^{-1}\mathbf{V}^t\boldsymbol{\Sigma}^{-1}\hat{\mathbf{A}}(\mathbf{x}). \quad (13)$$

This process of decomposing the high-dimensional attribute vector to low-dimensional attribute vector is called factor analysis. The $\mathbf{V}$-matrix obtained after decomposition contains the eigenvectors of largest l eigenvalues. These eigenvalues are from the Gaussian mixtures that model the attributes in the given clip. The computation complexity of calculating $\mathbf{r}$ is $O(QKl + Ql^2 + l^3)$.

3.4 Classification of Fine-grained Actions

Multi-class SVM is employed to classify the fine-grained actions and the obtained low-dimensional attribute vectors are used to find the similarity between two fine-grained actions. 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 (Hearst et al., 1998). The SVM is a supervised learning model, which minimizes the objective function

$$J = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i^T, \mathbf{x}) - \sum_{i=1}^{n}\alpha_i, \quad (14)$$

where α_i are lagrange's multipliers and n is the number of video clips. $K(\mathbf{x}_i^T, \mathbf{x})$ is the kernel function to obtain similarity between two vectors. During the testing process, the decision function for the low-dimensional test attribute vector $\mathbf{x}_t$ is given by

$$f(\mathbf{x}_t) = sign\left(\sum_{i=1}^{m}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_t) + b\right). \quad (15)$$

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

4 EXPERIMENTAL RESULTS

In the proposed method, a large GMM is trained on the HOF, MBH, and 3D-CNN descriptors separately for various mixtures ranging from 32, 64, 128, 256, and 512. The adapted means of the mixtures are concatenated resulting in a high dimension attribute vector. For example, an attribute vector obtained from a 32 component GMM is $(32 \times 108) = 3456$ where

(a) (b)

Figure 3: Confusion matrix of attribute vector for (a) JIGSAWS dataset and (b) KSCGR dataset.

32 is the number of mixtures and 108 is the dimension of the HOF feature descriptor. Also, as the components of GMM increase, the dimension of the attribute vector also increases. Also, it contains redundant attributes that may not contribute to a particular fine-grained action. So, the dimension of the attribute vector is reduced to 200-dimensions using the factor analysis method. The performance of the proposed approach is evaluated on the variety of fine-grained action datasets that are chosen from 2 different applications, namely, 'medical surgeries' and 'cooking'. In experiments, we extract features from final layer of pre-trained 3D-CNN network after fine-tuning on our datasets (Hara et al., 2018). Three independent GMMs are trained separately for HOF, MBH, and 3D-CNN descriptors to demonstrate the efficacy of the proposed approach. The detailed analysis of the performance of the proposed approach for each dataset is described in the following sub-sections.

4.1 JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS)

JIGSAWS dataset consists of videos recorded by endoscopic cameras placed on the right and left sides of a surgical robotic arm (Gao et al., 2014). Fine-grained actions, namely, '*suturing (SUT)*', '*needle passing (NP)*', and '*knot tying (KT)*' are performed by 8 different subjects. Each subject repeats all fine-grained action 5 times. The dataset contains 78 videos on suturing, 56 videos on needle passing, and 72 videos on knot tying. Figure 3a gives the confusion matrix of the best performance mixture model on JIGSAWS dataset. It can be seen from the figure that the total classification accuracy is close to the classification accuracy of each fine-grained action depicting that the proposed GMM captures attributes of all fine-grained actions uniformly. From the Figures 4a, 4b, 4c it can be observed the absolute discrimination of the three fine-grained actions.

Table 2 gives the comparison of the proposed approach with other deep learning baseline architectures on JIGSAWS dataset. Fawaz (Fawaz et al., 2018) leveraged the efficiency of CNNs to extract the patterns of motions performed in robotic surgery. The activation maps of CNNs highlight the parts which influence the classification of surgical tasks. Funke (Funke et al., 2019) investigated inflated 3D ConvNet to classify video snippets (few consecutive frames) extracted from untrimmed videos. It can be inferred from Table 2 that the proposed method performs on par with the existing supervised deep learning approaches.

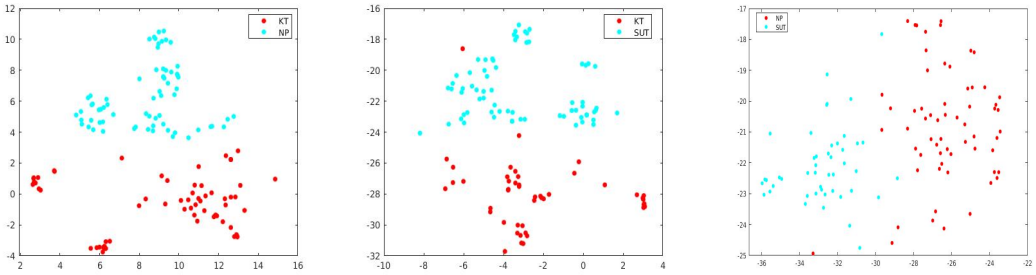
4.2 Max Planck Institute for Informatics (MPII cooking2)

MPII cooking2 dataset consists of cooking videos performed by 30 different subjects (Rohrbach et al., 2016). Each subject performs 62 different fine-grained actions such as '*cut dice*', '*cut stripes*', '*peeling*', etc. The dataset contains 273 untrimmed videos, where train and test sets are split based on the number of subjects. Train set consists of videos performed by 20 subjects and remaining 10 subjects are considered for test set. The class-wise classification performance on MPII cooking2 dataset is shown in Figure 6. It compares the correctly classified samples (in blue colour) with the total number of test samples (in green colour). The row at the bottom gives the class-wise classification accuracy. From the figure, it can be observed that the proposed approach differentiates well among the fine-grained actions such as '*take out from cupboard*', '*take out from drawer*', '*take out from fridge*' etc, without the need for explicit object detection. Figure 5a shows the clear discrimination of two most confusing fine-grained actions such as '*take out from cupboard*' and '*take out from drawer*'. It can be observed from Figures 4, 5, that proposed model can model the subtle interactions between human and object efficiently.

Table 3 compares the performance of the proposed approach with existing methods. The pose-based approach gives low performance because this framework is based on the trajectories extracted from the joints, which are noisy. The dense trajectories approach performs better than pose-based approach because of capturing the robust motion information.

Table 1: Classification accuracy (%) of SVM classifier on various number of mixtures. BTL, ATL refer to before transfer learning and after transfer learning, respectively.

	JIGSAWS			KSCGR (BTL)			KSCGR (ATL)			MPII cooking2		
	HOF	**MBH**	**3DCNN**	**HOF**	**MBH**	**3DCNN**	**HOF**	**MBH**	**3DCNN**	**HOF**	**MBH**	**3DCNN**
32	84.5	97	96	61.5	66.7	62	76.2	80.6	78.4	63.7	72.1	69.6
64	85.9	97.6	96.3	65.9	66.7	64	77.4	81.7	78.7	65.4	72.9	70.1
128	87.9	**99.2**	97.1	63.1	**69**	65.3	78.6	**83.7**	80.5	66.7	74.5	72.4
256	90.3	98.5	96.9	58.3	64.3	63.9	77.8	81.7	81.2	68.8	**75.7**	72.9
512	93.7	98.5	96.7	57.9	63.1	62.8	80.2	80.9	80.2	68	74.7	71.8



(a) (b) (c)

Figure 4: t-SNE plot of attribute vectors for (a) *knot tying (KT)* vs *needle passing (NP)* (b) *knot tying (KT)* vs *suturing (SUT)* (c) *needle passing (NP)* vs *suturing (SUT)*. Here, the axes represent first and second dimensions in the factor analysis

Table 2: Performance (%) comparison on JIGSAWS dataset.

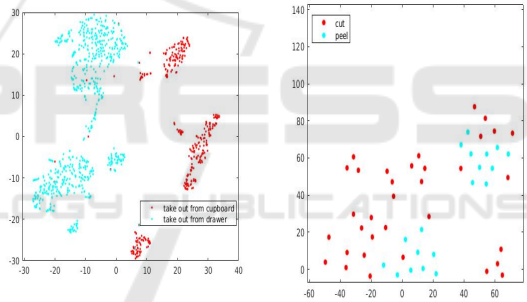
Method	**Accuracy (%)**
vector space model (Forestier et al., 2017)	82.36
convnet (Wang and Fey, 2018)	93.06
CNN (Fawaz et al., 2018)	97.3
3D Conv Net (Funke et al., 2019)	98.3
TSN (Wang et al., 2016)	98.33
TSM (Lin et al., 2019)	99.1
Proposed approach	**99.2**

Table 3: Performance (%) comparison on MPII cooking2 dataset.

Method	**mAP (%)**
Pose-based approach (Rohrbach et al., 2016)	24.1
Hand-cSIFT + Hand-Trajectories (Rohrbach et al., 2016)	43.5
Dense trajectories (Rohrbach et al., 2016)	44.5
Region-sequence CNN (M. et al., 2018)	70.3
TSN (Wang et al., 2016)	68.5
TSM (Lin et al., 2019)	71.2
Proposed approach	**73.7**

4.3 Kitchen Scene Context-based Gesture Recognition (KSCGR)

KSCGR dataset consists of cooking videos, containing fine-grained actions performed by 5 different subjects to assess the various human gestures (A et al., 2013). There are 8 fine-grained actions, namely, '*break*', '*mix*', '*bake*', '*turn*', '*cut*', '*boil*', '*season*', and '*peel*'. The dataset contains 25 training videos



(a) *take out from cupboard* vs *take out from drawer* (b) *cut* vs *peel*

Figure 5: t-SNE plot of attribute vectors for (a) MPII cooking2 dataset (b) KSCGR dataset.

and 10 testing videos each ranging from 5 to 10 minutes long. The fine-grained actions such as 'boil', 'bake', and 'peel' are hard to recognize as there is no salient motion present in such actions. Also, the KSCGR dataset has only a few number of training videos, therefore GMM is unable to learn the distribution of data efficiently. So, the total variability matrix and GMM trained on MPII cooking2 dataset are used to form the attribute vectors for fine-grained action videos in the KSCGR dataset. The reason for using the model trained on MPII cooking2 dataset is that it consists of fine-grained cooking actions similar to that of KSCGR dataset, namely, 'peel', 'mix', 'cut' etc and the dataset contains enough data to train the GMM well. The classification accuracy on the

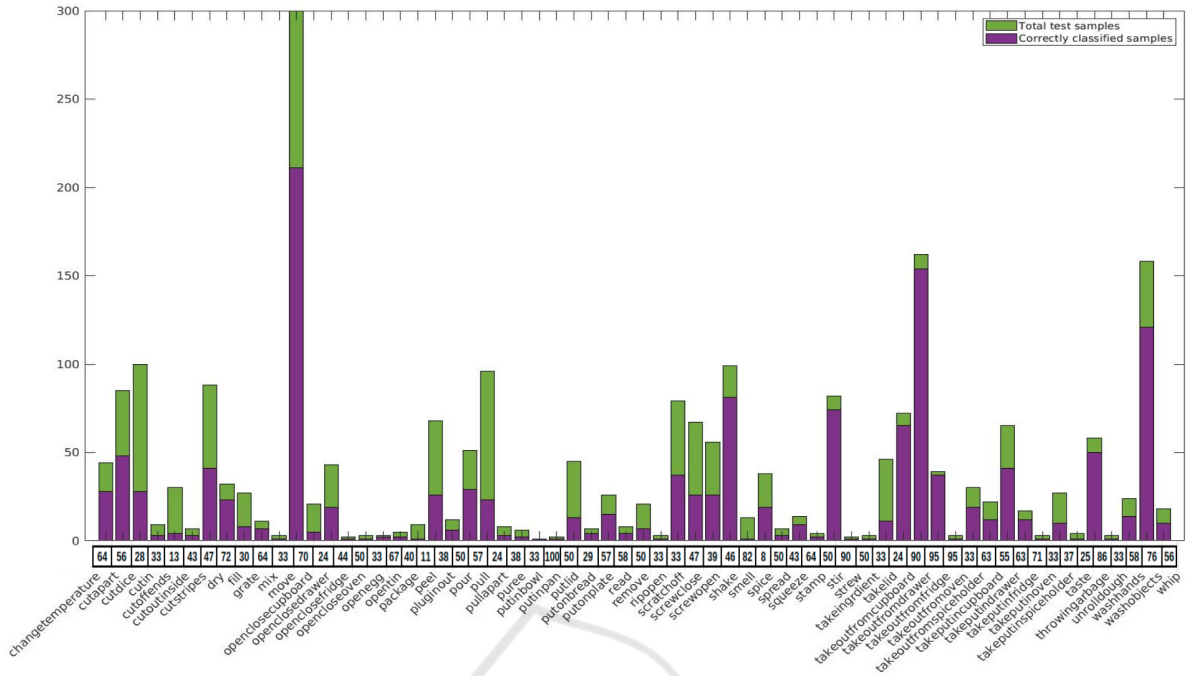


Figure 6: Class-wise classification performance on MPII cooking dataset (best viewed in color). The row at bottom gives the class-wise classification accuracy.

KSCGR dataset after transfer learning outperforms the result obtained for GMM trained only on KSCGR dataset as shown in Table 1. This shows that the proposed approach is able to model the attributes better where multiple fine-grained actions share the common attributes. The performance of each class is presented in the form of the confusion matrix as shown in Figure 3b. It can be observed that the fine-grained actions such as 'cut' and 'peel' are misclassified because of the overlap of the attribute vectors of these two actions as shown in Figure 5b.

Table 4 gives the performance comparison of proposed approach with existing methods on KSCGR dataset. Ni (Ni et al., 2016) leveraged the low-level features by encoding the IDT features using Fisher vectors in order to classify the fine-grained actions using SVM. Granada (Granada et al., 2017) proposed deep neural architecture for recognising kitchen activities using ensemble of machine learning models and hand crafted features to extract efficient representation of data. It can be observed from Table 4 that the proposed approach performs better than the existing deep learning architectures with large margin.

Table 4: Performance comparison on KSCGR dataset.

Method	F-score
IDT-IFV-SVM (Ni et al., 2014)	0.76
TSN (Wang et al., 2016)	0.65
TSM (Lin et al., 2019)	0.69
RGB + OF + CNN + SVM (Granada et al., 2017)	0.70
RGB + OF + CNN + NN (Granada et al., 2017)	0.72
Proposed approach (after transfer learning)	**0.824**

4.4 Ablation Study

The 128 mixture GMM trained on MBH descriptors gives the best performance on both JIGSAWS and KSCGR datasets (shown in Table 1) as these datasets contains few actions. But for MPII cooking2 dataset, 256 mixture GMM performs better than 128 mixtures, as 128 mixtures may not be enough to model all 62 fine-grained actions. Also, as the number of mixtures increases, the classification accuracy reduces. This is due to the fact that the GMM needs more local information in order to capture the attributes of fine-grained actions.The GMM trained on MBH descriptors performs better than 3D-CNN features as MBH captures local motion information effectively. In experiments, we evaluate the recent state-of-the-art approaches such as TSN (Wang et al., 2016), TSM (Lin et al., 2019) on three fine-grained datasets. The proposed approach performs better than the existing deep learning approaches. This is because the deep learning methods fail to generalise on smaller datasets, whereas our proposed GMM model is able to capture the attributes that multiple fine-grained actions share relatively better.

5 CONCLUSION

In this paper, a framework is proposed to learn an efficient low-dimensional representation of fine-grained actions. The fixed dimensional attribute vector performs on-par when compared with the other supervised techniques on JIGSAWS, KSCGR, and MPII cooking2 datasets. The effectiveness of the attribute vector for classification on the KSCGR dataset proves that the proposed method performs better even when there is a few number of samples for each action. We demonstrate the generalization of the proposed approach by evaluating on a wide variety of fine-grained action datasets. Also, the proposed approach can be adapted in applications such as medical, elderly assistance, autonomous vehicles etc.

REFERENCES

A, S., K, K., D, D., G, M., and H, S. (2013). Kitchen scene context based gesture recognition: A contest in icpr2012. *International Workshop on Depth Image Analysis and Applications*, 7854:168–185.

Alexandros, I., Anastastios, T., and Ioannis, P. (2014). Discriminant bag of words based representation for human action recognition. *Pattern Recognition Letters*, 49:185–192.

Andrej, K., George, T., Sanketh, S., Thomas, L., Rahul, S., and Li, F.-F. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Cheng, M., Zhang, Z., Lin, W., and Torr, P. (2014). Bing: Binarized normed gradients for objectness estimation at 300fps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3286–3293.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. (2018). Evaluating surgical skills from kinematic data using convolutional neural networks. *CoRR*, abs/1806.02750.

Forestier, G., Petitjean, F., Senin, P., Despinoy, F., and Jannin, P. (2017). Discovering discriminative and interpretable patterns for surgical motion analysis. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 136–145. Springer.

Funke, I., Mees, S. T., Weitz, J., and Speidel, S. (2019). Video-based surgical skill assessment using 3d convolutional neural networks. *CoRR*, abs/1903.02306.

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D. D., et al. (2014). Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Miccai workshop: M2cai*, volume 3, page 3.

Granada, R. L., Monteiro, J., Barros, R. C., and Meneguzzi, F. R. (2017). A deep neural architecture for kitchen activity recognition. In *The Thirtieth International Flairs Conference*.

Hao, Y., Chunfeng, Y., Bing, L., Yang, D., Junliang, X., Weiming, H., and J, M. S. (2019). Asymmetric 3d convolutional neural networks for action recognition. *Pattern recognition*, 85:1–12.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Heeseung, K., Yeonho, K., S, L. J., and Minsu, C. (2018). First person action recognition via two-stream convnet with long-term fusion pooling. *Pattern Recognition Letters*, 112:161–167.

Herve, J., Florent, P., Matthijs, D., Jorge, S., Patrick, P., and Cordelia, S. (2011). Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716.

Ivan, L. (2005). On space-time interest points. *International journal of computer vision*, 64(2-3):107–123.

Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093.

M., M., N., M., Y., L., A., L., and R, S. (2018). Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521.

Manel, S., Mahmoud, M., and Ben, A. C. (2015). Human action recognition based on multi-layer fisher vector encoding method. *Pattern Recognition Letters*, 65:37–43.

Maria, C. J. and Joan, C. (2018). Human action recognition by means of subtensor projections and dense trajectories. *Pattern Recognition*, 81:443–455.

Miao, M., Naresh, M., Yibin, L., Ales, L., and Rustam, S. (2018). Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521.

Ni, B., Paramathayalan, V. R., and Moulin, P. (2014). Multiple granularity analysis for fine-grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 756–763.

Ni, B., Yang, X., and Gao, S. (2016). Progressively parsing interactional objects for fine grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1020–1028.

Reynolds, D., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. In *Digital Signal Process.*, volume 10, pages 19–41.

Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., and Schiele, B. (2016). Recognizing fine-grained and composite activities using

hand-centric features and script data. *International Journal of Computer vision (IJCV)*, 119(3):346–373.

Singh, B., Marks, T. K., Jones, M., Tuzel, O., and Shao, M. (2016). A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1970.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Wang, H., Klaser, A., Schmid, C., and Liu, C. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Wang, Z. and Fey, A. M. (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International journal of computer assisted radiology and surgery*, 13(12):1959–1970.

Yamin, H., Peng, Z., Tao, Z., Wei, H., and Yanning, Z. (2018). Going deeper with two-stream convnets for action recognition in video surveillance. *Pattern Recognition Letters*, 107:83–90.

Zhigang, T., Wei, X., Qianqing, Q., Ronald, P., C, V. R., Baoxin, L., and Junsong, Y. (2018). Multi-stream cnn: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, 79:32–43.

Zhou, Y., Ni, B., Hong, R., Wang, M., and Tian, Q. (2015). Interaction part mining: A mid-level approach for fine-grained action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3323–3331.