Multimedia Tools and Applications (2021) 80:24303–24318
https://doi.org/10.1007/s11042-021-10721-6

A statistical framework for few-shot action recognition

Mark Haddad[1] · Vahid K. Ghassab[1] · Fatma Najar[1] · Nizar Bouguila[1]

Received: 2 July 2020 / Revised: 23 October 2020 / Accepted: 10 February 2021 /
Published online: 5 April 2021

Abstract

Along with the exponential growth of online video creation platforms such as Tik Tok and Instagram, state of the art research involving quick and effective action/gesture recognition remains crucial. This work addresses the challenge of classifying short video clips, using a domain-specific feature design approach, capable of performing significantly well using as little as one training example per action. The method is based on Gunner Farneback's dense optical flow (GF-OF) estimation strategy, Gaussian mixture models, and information divergence. We first aim to obtain accurate representations of the human movements/actions by clustering the results given by GF-OF using K-means method of vector quantization. We then proceed by representing the result of one instance of each action by a Gaussian mixture model. Furthermore, using Kullback-Leibler divergence (KL-divergence), we attempt to find similarities between the trained actions and the ones in the test videos. Classification is done by matching each test video to the trained action with the highest similarity (a.k.a lowest KL-divergence). We have performed experiments on the KTH and Weizmann Human Action datasets using One-Shot and K-Shot learning approaches, and the results reveal the discriminative nature of our proposed methodology in comparison with state-of-the-art techniques.

Keywords Action recognition · One-shot learning · K-shot learning · Mixture models · Information divergence · Action representation

✉ Mark Haddad
mar_had@encs.concordia.ca

Vahid K. Ghassab
vahid.khorasani@concordia.ca

Fatma Najar
f_najar@encs.concordia.ca

Nizar Bouguila
nizar.bouguila@concordia.ca

[1] Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal QC, Canada

1 Introduction

In this paper, we propose a novel action recognition framework whose goal is to classify short action video clips to their respective actions by automatically matching their representations to trained ones. The trained representations are essentially labeled instances of each action, as shown in the upper left of Fig. 1, that are used in a few-shot learning setting to achieve a few-shot action recognition task. The framework is flexible enough to be extended in various ways according to the application and could for example be integrated in users' devices to classify their videos using little training data.

Requests for new action (e.g. dance) challenges are emerging on a daily basis, and our framework is designed to be able to effectively learn each new action using as little as one instance of it, and classify new videos using the learned instances. An overview of the process flow is displayed in Fig. 2 and goes as follows: Initially, the input dataset is split into training and testing sets that both go through the same feature extraction process. This process initially tracks the actors in the videos and places a bounding box around them, computes the dense optical flow inside the box, and clusters the optical flow vectors using the KMeans algorithm. Subsequently, classification is achieved using a similarity check method which employs Gaussian Mixture Models and Kullback-Leibler divergence between the KMeans clusters of the training and testing videos. A visual representation of the similarity measurement is demonstrated in Fig. 1, in which the KMeans cluster centers of a trained "waving" action are used in an attempt to find similar movement patterns in a test video.

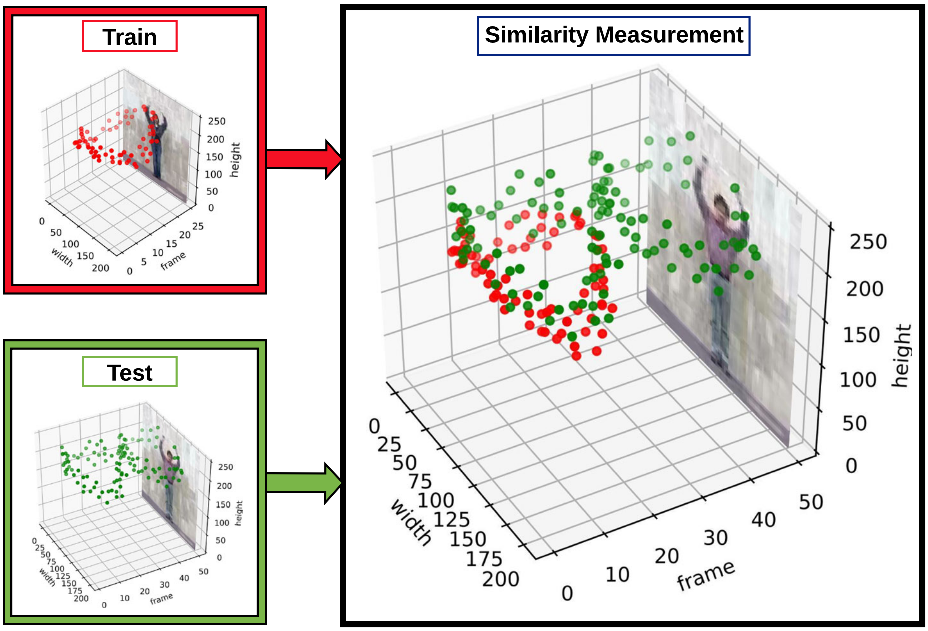

Fig. 1 Similarity between clustered optical flow centers (KMeans) of a trained "Two-hands wave" action (Red) and a test one (Green) from the Weizmann Human Action dataset

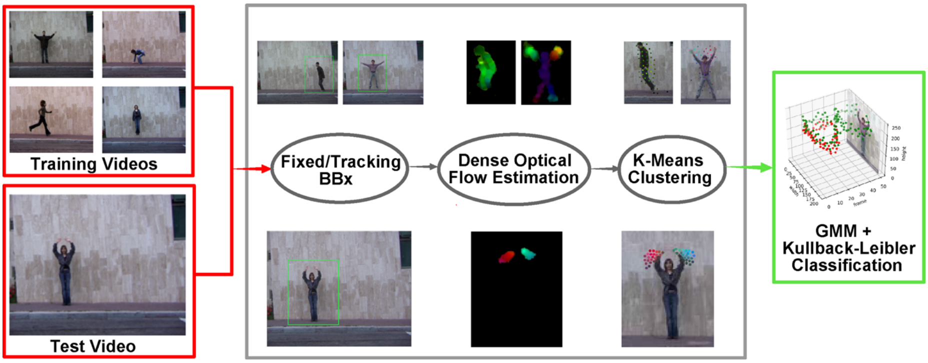
Training Videos
Test Video
Fixed/Tracking BBx
Dense Optical Flow Estimation
K-Means Clustering
GMM + Kullback-Leibler Classification

Fig. 2 Overview of the process flow. The training and testing videos (shown in red) are used as inputs and go through a feature extraction process (shown in grey), followed by a similarity check process (shown in green) in which classification is achieved

1.1 Data preprocessing

The goal is to pre-process all the data in a way that will maximize the quality of the extracted features, while discarding unmeaningful ones. To start off, a fixed size bounding box *(BBx)* was employed around the actors in each video to maximize classification accuracy while minimizing background noise, however this step remains optional. Resizing was then done to each frame, and area interpolation was used to cover up for lost information. For actions that involve significant lateral movement, a high-speed tracking method based on kernelized correlation filters [18] was employed on the actors to automatically keep the *BBx* around them.

1.2 Feature extraction

Following the preprocessing step in which each actor was tracked and surrounded by a *BBx*, feature extraction was completed using Gunner Farneback's optical flow [9, 10]. The feature, being the optical flow in its raw 2D format, was extracted by computing the optical flow between each two consecutive frames t and $t+1$. The result was essentially a set of 2D vectors (u, v), each representing the horizontal and vertical movement of a pixel at position (x, y), respectively. Noise was reduced by thresholding vectors under a certain magnitude. Although the application of a threshold generally positively impacts the classification results, the value of the threshold itself was not critically important due to the flexibility of the proposed framework. The 2D optical flow feature was then converted to a 5D set of vectors (x, y, u, v, t) by concatenating the position of each pixel in 3D space to it; (x, y, t) represent the horizontal position, vertical position, and frame number of the pixel in 3D space, respectively. Finally, K-means clustering was applied on k frames long sub-clips, resulting in clusters of similar optical flow vectors. The center of each cluster was a 5D vector (x, y, u, v, t) representing its mean position and optical flow values.

1.3 Paper organization

The rest of the paper is arranged as follows. The related works, including their novelties and drawbacks, are illustrated in Section 2 in addition to our contributions. Section 3 describes the methodology and the mathematical background behind the proposed framework. The different experiments that have been carried out are detailed and empirically examined in Section 4. Limitations and potential improvements in future works are presented in Section 5, and finally, the paper is concluded in Section 6.

2 Related work

Recognition is a field that concentrates on a classical problem in computer vision, which is determining whether the information on images or video frames contains a specific feature, object, or activity. Such field includes "object recognition", "Human action recognition", "identification" and "detection" [5, 6, 38–42, 44]. On the other hand, applying deep learning, image restoration and classification has facilitated the study over different sub-branches of recognition. Some of the novel deep learning, image restoration and classification applications which are recognized for this purpose can be found in [11, 12, 16, 17, 22, 24, 37, 43].

The "Human action recognition" field is an active and important area in computer vision. Related comprehensive research works can be found in [3, 21, 34, 45]. In this regard, spatiotemporal interest points and feature descriptors for human action recognition have been researched in [1, 19, 26], which include a wide range of methodologies and described as "Bag of visual and video words". The strength of these methodologies is their robustness to occlusion, whereas their drawback is their locality and distribution of content understanding, and their sensitivity to several intermediate processes such as classifiers. There is another set of techniques that focuses on detecting a bounding box, which includes the person executing the action. These methods include spatiotemporal shapes using contours for body tracking [35], spatiotemporal volumes using silhouette images [30] and space-time gestures [8]. Such methodologies ignore the primitive human sub-actions, which are considered a drawback for their representations.

There is another set of methodologies which considers the location knowledge or body parts appearances. For instance, landmark trajectory features of body parts have been researched in [36]. Additionally, the learning of cascade of filters has been proposed by Ke et al. [20] for accurate spatiotemporal localization and detection purposes. Such approaches are challenging issues in the field of human action recognition since a completely supervised strategy is not ensured.

The problem of long-term visual tracking has been addressed in [4] where ascribable to deformation, abrupt motion, heavy occlusion and out-of-view, the target objects undergo significant appearance variation. Accordingly, the task of tracking into translation and scale estimation of objects has been decomposed. In another work, an adaptive region proposal scheme with feature channel regularization has been provided for facilitating robust object tracking [23]. Correspondingly, the unsupervised video object segmentation task has been addressed in [32] where the method was denominated as CO-attention Siamese Network (COSNet). Recently, another video object segmentation (VOS) work has been proposed which unlike most existing methods which rely heavily on extensive annotated data, this method addresses object pattern learning from unlabeled videos [33].

Contributions Despite the significant progress which has formely been performed, there are several challenges in the field of human action recognition. For instance, the variation of the camera position relative to the subject may create confusions in the human action detection and classification. Moreover, similarities in different action categories may cause action misclassifications. In this work, we have tried to overcome such challenges by presenting a human action representation and classification framework that automatically matches human action test videos to trained ones. The action representation is based off the repetitive nature of human actions, and can be utilized effectively in one-shot or k-shot learning settings [13, 25]. The importance of our contributions can be described as follows:

- Classifying human actions by automatically matching their representations to trained ones in videos.
- Representing human actions considering their repetitive nature.

3 Methodology

In this section, we explain the concepts and fundamental strategies behind our proposed structure. This includes the optical flow application that has been implemented, followed by the action instance representation that we have proposed and the approach we have used to obtain meaningful similarity measurements between different actions for the classification process.

3.1 Parameterized displacement fields

In this subsection, we are going to describe the parametrized displacement fields which we have applied in Farneback optical flow estimation considering two consecutive video frames using the eight-parameter model in a two dimensional space [9, 10]. For this purpose, we define the global parameterized displacements considering polynomials which represent the neighborhood of a pixel in each of our video frames as follows

$$\begin{aligned} d_x(x,y) &= a_1 + a_2x + a_3y + a_7x^2 + a_8xy, \\ d_y(x,y) &= a_4 + a_5x + a_6y + a_7xy + a_8y^2, \end{aligned} \tag{1}$$

where x and y are the horizontal and vertical coordinates of corresponding pixels in two consecutive video frames; and d_x and d_y illustrate the parametrized displacement polynomial with respect to x and y. Furthermore, $a_1, a_2, \cdots, a_8$ are expansion coefficients considering the polynomial expansions of both video frames. Equation 1 can be rewritten as

$$D = PS, \tag{2}$$

$$P = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{pmatrix}, \tag{3}$$

$$S = (a_1 \; a_2 \; a_3 \; a_4 \; a_5 \; a_6 \; a_7 \; a_8)^T, \tag{4}$$

where $D = \langle d_x, d_y \rangle$ is the global displacement, and P and S stand for the polynomial matrix and the solution, respectively. In addition, the polynomial expansion is the neighborhood approximation of each pixel with a polynomial. Accordingly, the quadratic polynomial in a local coordinate system can be represented as

$$F(X) \sim X^T AX + BX + C, \tag{5}$$

where $X = \langle x, y \rangle$ is a pixel vector considering its direction in the video frame and F is the polynomial expansion of pixels neighborhood in the video frame. Furthermore, A, B, and C are the coefficients of such neighborhood polynomial expansion where A represents a symmetric matrix, B is a vector and C is considered a scalar. By applying the global displacement in (5), we end up with

$$\begin{aligned} F(X-D) &\sim (X-D)^T A(X-D) + B^T(X-D) + C \\ &= X^T AX + (B-2AD)^T X + D^T AD - B^T D + C. \end{aligned} \tag{6}$$

Accordingly, by defining $B' = B - 2AD$ and $\Delta B = \frac{B'-B}{2}$ and considering (3), we minimize the following weighted least square problem for calculating our desired solution

$$\sum_j \omega_j \left\| A_j P_j S - \Delta B_j \right\|, \tag{7}$$

where j is the pixel index and ω_j represents the weight of the corresponding pixel. Therefore, the solution is calculated as follows

$$S = \left(\sum_j \omega_j {P_j}^T {A_j}^T A_j P_j \right)^{-1} \sum_j \omega_j {P_j}^T {A_j}^T \Delta B_j. \tag{8}$$

The application of the parametrized displacement fields solution displayed in (8) is illustrated on the left side of Fig. 3, in which a “bending” action from the Weizmann Human Action dataset was used as an input. Furthermore, the right side is the result of clustering of the optical flow points using K-means. In Section 3.2 we will describe how this feature extraction method may be utilized on a spatiotemporal level to obtain a valuable representation tool for human action classification.

3.2 Proposed framework

As seen in Fig. 4, obtaining K-means clusters gives an accurate 3D representation of the movement in a video. The idea behind our approach involves the use of just one (or k) example(s) of each action in the training phase. Since our work is instance-based oriented, we focus on obtaining a representation of a single instance/repetition of each action. In videos which contain many repetitions of the same action, we only focus on one of the occurrences, typically the one that looks the most representative or general for the action.

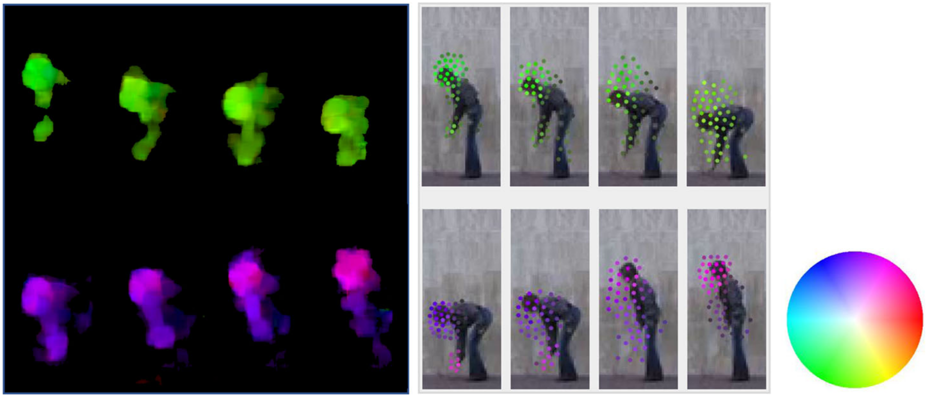

Fig. 3 Application of dense optical flow (on left) and K-means clustering (on right) on a “Bending” action from the Weizmann Human Action dataset. Colors correspond to the flow magnitude and direction, as per the color wheel

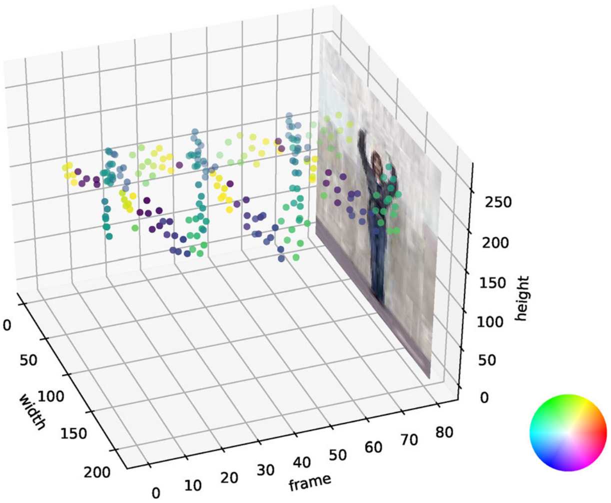

Fig. 4 3D Representation of K-means clusters of optical flow for the "Wave 2" action of Daria from Weizmann Human Action dataset. Colors correspond to the mean flow magnitude and direction of each cluster, as per the color wheel. Three full repetitions of the action are clearly discernable

Once an instance for each action is obtained during the training phase, we employ the following method to compare those instances to groups of K-means points found in test videos. In this regard, we propose that the K-means clusters of each action instance be modeled by a mixture of Gaussian distributions, resulting in a set of Gaussian components defined as follows

$$p(x|\Theta) = \sum_{j=1}^{M} p_j \mathcal{N}\big(x; \mu_j; \Sigma_j\big)\,, \tag{9}$$

where p_j is the mixing parameter of component j $\left(0 \le p_j \le 1, \sum_{j=1}^{M} p_j = 1\right)$, Θ is the set of all the parameters $(p_1, \ldots, p_M, \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M)$ and $\mathcal{N}\big(x; \mu_j; \Sigma_j\big)$ is the j-th Gaussian distribution given by the mean μ_j and the covariance matrix parameter Σ_j

$$\mathcal{N}\big(x; \mu_j; \Sigma_j\big) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_j|^{1/2}} \times \exp\left\{-\frac{1}{2}\left(x - \mu_j\right)^T \Sigma_j^{-1}\left(x - \mu_j\right)\right\}. \tag{10}$$

Each Gaussian component represents part of an action, meaning that the mean μ of each component is a 5D vector (x, y, u, v, t) representing the position (x, y, t) and magnitude (u, v) of a group of K-means clusters that constitute a lower-level action, or sub-action (e.g. left arm moving up, right leg moving to the left).

The parameters of each Gaussian mixture model are then estimated using the Expectation-Maximization estimation algorithm (EM) where the log-likelihood is derived with respect to the mean, the covariance matrix, and the mixing weight. Starting with the expected value of the posterior probabilities, all the parameters are updated until convergence of the likelihood. Subsequently, Kullback-Leibler (KL) divergence measure is employed in an attempt to find similarities between the GMM representations of the trained action instances, and the ones being generated in different sections of each test video.

Considering the single Gaussians $p(x) = \mathcal{N}\left(x; \mu_p; \Sigma_p\right)$ and $q(x) = \mathcal{N}\left(x; \mu_q; \Sigma_q\right)$, the KL-divergence is represented as follows [15]

$$KL_{GMM}\left(p||q\right) = \frac{1}{2}\left[\log\frac{\left|\Sigma_p\right|}{\left|\Sigma_q\right|} + \mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) - k + \left(\mu_q - \mu_p\right)^T \Sigma_q^{-1}\left(\mu_q - \mu_p\right)\right], \quad (11)$$

where k is the dimension of both distributions, μ_p and μ_q stand for the mean values of the Gaussians, Σ_p and Σ_q represent the covariance values and $\mathrm{tr}(\cdot)$ the trace of a matrix.

In order to compute the KL-divergence between two GMMs, we consider the approximation proposed by Goldeberger et al. [14] as follows

$$KL_{GMM}\left(f||g\right) = \sum_{i=1}^{m}\omega_{f,i}\left(KL_G\left(f_i||g_{\pi(i)}\right) + \log\frac{\omega_{f,i}}{\omega_{g,\pi(i)}}\right), \quad (12)$$

where $\pi(i) = \arg\min_j \left(KL_G\left(f_i||g_j\right) - \log\omega_{g,j}\right)$, f and g are two GMMs including f_i and g_i for $i \in \{1, \cdots, m\}$ as their Gaussian distributions. Moreover, $\omega_{f,i}$ and $\omega_{g,i}$ are the corresponding weights and m is the total number of Gaussian components.

Since each trained action has its own GMM representation, the classification process is done by matching each test video to the trained action with which the KL-divergence value was the lowest. The proposed classification framework is described in Algorithm 1.

Algorithm 1 The proposed classification framework using similarity measurement.

```
function Classification ($KM, n, GMM$);
Input  : KMeans parameters $KM$ of test videos $\mathcal{X}$
         Number of frames $n$ of training $block_t$
         Gaussian Mixture Model parameters $GMM$ of training $block_t$
Output: Classification labels $t$ of videos $\mathcal{X}$
foreach Video $X_i$ in $\mathcal{X}$ do
    foreach $block_j$ in the set of $blocks_n$ do
        $GMM_j = \mathrm{GMM}(block_j)$, (9);
        $KL_j = KL\left(block_j, block_t\right)$, (12) ;
    end
    $min_i = \mathrm{Min}\ KL_j$
    Label $X_i$ as $t$ where $min_i = KL\left(block_j, block_t\right)$
end
return labels $t$
```

4 Experiments

The conducted experiments involved training using a set of actions and classifying test ones using one-shot and k-shot learning approaches. Trials were conducted using several assumptions in an attempt to increase the representation quality of each action, hence maximizing classification accuracy.

4.1 Dataset

Our experiments were conducted on the Weizmann Human Action [2] and the KTH [7] datasets, which contain actions that resemble the ones seen on online platforms (Short, contain one or more repetitions of the action and recorded using a fixed camera). The Weizmann dataset contains 90 low-resolution videos, consisting of 10 natural actions (bend, jumping jack, jump forward, jump in place, gallop sideways, run, skip, walk, wave one hand, wave two hands). As for the KTH dataset, it contains 600 action videos, involving 25 subjects performing 6 different actions (walking, jogging, running, boxing, hand waving, hand clapping) in 4 dissimilar scenarios. As mentioned in Section 1.1, the videos were preprocessed prior to feature extraction. Actions in which the actors are not horizontally moving were placed in a bounding box and upscaled three times, whereas the rest of the actions went through an additional process, described earlier as *Tracking BBx*, in which the actors were automatically tracked to keep them centered inside the *BBx*.

4.2 Human action classification

4.2.1 Training

Our goal was to obtain a representation of a single instance for each action. In this regard, after computing GF-OF between two consecutive frames, K-means was applied using 50 components ($K = 50$) on sub-clips $[f, f + 2]$ consisting of three consecutive frames. A moderately larger number of frames per sub-clip may have also been employed for faster computation, without having a considerable impact on classification results. Sub-clips in which little to no movement was present (e.g. transition from bending down to going back up) employed a lower value of K clusters, to prevent them from holding or being concentrated upon few optical flow points.

During the training process for one-shot learning, one video of each action was used (k videos were used for k-shot learning). Training videos in which the action was only executed once had all their K-means clusters, which represent that single instance, gathered. On the other hand, training videos in which more than one repetition of the action was completed, had the K-means clusters of only one of those instances saved. This process was completed by setting a range of frames which contain only one instance for each action. The length f (number of frames) of each action occurrence was also stored and used in the training process described in Section 4.2.2. Finally, the K-means clusters of each action were represented by a Gaussian mixture model consisting of n Gaussian components.

Gaussian components Experiments were conducted to determine the ideal number n of Gaussian components per mixture. The results illustrated in Fig. 5 demonstrate that the highest average classification accuracies are achieved when $n = 9$. This signifies that 9 Gaussian components are sufficient to accurately represent a fully executed action.

Additionally, Gaussian components of similar action instances may resemble each other in the (x, y, u, v) dimensions, however, they often do not occur in the same range on the temporal axis (t). To deal with this drawback, we employed a simple method in which the Gaussian mixture of each action instance was assumed to begin at $t = 0$ on the temporal axis. This was done by subtracting the lowest t value of all the K-means points, which are within the range of frames of interest, from the t values of all the other points within that same range. This process was employed prior to each Gaussian mixture generation in both training and testing steps. For example, selecting the second "Wave2" instance displayed in Fig. 4 for training would involve the employment of this process to the K-means points within the range of that instance so that they end up being within the frame range $t \in [0, 25]$ instead of $t \in [30, 55]$.

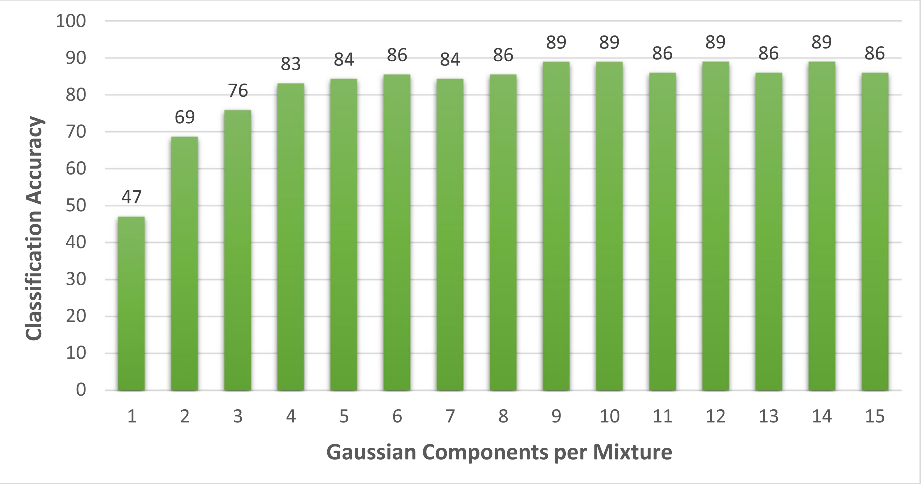

Fig. 5 Classification accuracy with different numbers of Gaussian components n per mixture using the Weizmann dataset. The highest accuracies were achieved when at least 9 components were used ($n = 9$)

4.2.2 Testing

Once one, or k, representations of each action were carried out, the testing process went as follows: Each video in the dataset, except the ones used in training, went through a similarity measurement process in which an attempt to find similarities between the trained actions and the data of the test video was completed. This was done by calculating the KL-divergence between the Gaussian mixture of each trained action and different Gaussian mixtures in the testing video. Those Gaussian mixtures were generated on different f frames long blocks. One of the assumptions we made was that all similar actions have instances executed over the same number of frames f. For instance, a trained "bending" action consisting of 50 total frames, had an f value set to 50. Therefore, when an attempt was made to find similarities between this "bending" action and the action in the test video, 50 frames long blocks of K-means points were represented as Gaussian mixtures. After each cluster was generated, KL-Divergence was applied to obtain a measure of similarity between the training and the generated testing probability distributions. Finally, after obtaining a measure of similarity using each trained action, the test video was classified by matching it to the trained video with which the KL-divergence value was the lowest. All classification accuracies demonstrated are averages of 5 runs.

One-shot learning In some cases, two instances of a same action executed by two different actors have a significant resemblance between each other from a temporal perspective. An example of such case is demonstrated in Fig. 1, in which the actors "Daria" and "Denis" from the Weizmann dataset have very similar "Two-hands wave" actions. However, in general cases, a clear difference was noticed in the number of frames f required to represent one same action. This is due to the presence of variance in the execution time of an action from person to person. Due to this observation, we conducted some experiments, using one-shot learning, to check how our assumption regarding fixing the value of f according to the training data would affect the classification results. The experiments involved replacing f by $f + \Delta$ with $\Delta \in [1, 15]$, in which Δ represents a fixed number of additional frames

Table 1 Classification accuracies for one-shot learning of proposed work and similar works using Weizmann dataset

Seo and Milanfar [29]	75%
Yang [34]	80%
FSHMM [28]	81.5%
MAP+SHMM [27]	81.88%
MAP+SHMM (Relaxed) [27]	87.12%
Proposed	**89.4%**

The accuracy of our work are highlighted in bold

ranging from 1 to 15. The results showed only a slight fluctuation of ±2% in accuracy as Δ increased, confirming the validity of our assumption.

Experiments conducted using one-shot learning lead to an average classification accuracy of **89.4%** for the Weizmann dataset and **73.1%** for the KTH one. The accuracies of our work have been compared to other works which implemented one-shot/k-shot learning frameworks on the same datasets. The results displayed in Tables 1 and 2 show that our work has the highest accuracy compared to other works. The confusion matrix in Fig. 6, shows that when only one example per action is used during training in the Weizmann dataset, the main source of misclassification happens in the "skip" action, in which 50% of the test videos were wrongly classified, and received prediction labels of "side" or "walk" instead. In the case of the KTH dataset, the main source of misclassification was between the "jogging" and "running" actions, which are highly similar in nature. A solution we will be implementing in an attempt to fix such problem in future works is to automate the hyperparameters adjustment, as discussed in Section 5.

Table 3 demonstrates the KL-Divergence values between a set of trained actions using one-shot learning and the ones of actor "Daria" from the Weizmann Human Action dataset. The classification process was done by matching each action executed by "Daria" to the trained one with which the KL-Divergence value is the lowest. It is perceivable, that actions which share some similarities with each other, have lower divergence values between each other compared to ones that do not share ample similitude.

k-shot learning Following one-shot learning, each experiment involved incrementally training *one* additional example of each action prior to going through the classification process. Each additional action video used in training was removed from the test dataset. The results shown in Fig. 7 compare our classification accuracies using different values of k with different works. The graph shows that using as little as one training example per action (k = 1), a classification accuracies of **73.1%** and **89.4%** were achieved for the KTH and Weizmann datasets, respectively, compared to 80% for Yang [34], and 30% for BoVW [31]. As the number of training examples k increases, the classification accuracies can be seen to increase the most significantly when k is low ($k < 4$).

Table 2 Classification accuracies for one-shot learning of proposed work and similar works using KTH dataset

Seo and Milanfar [29]	65%
SHMM [27]	70.4%
FSHMM [28]	71.8%
Proposed	**73.1%**

The accuracy of our work are highlighted in bold

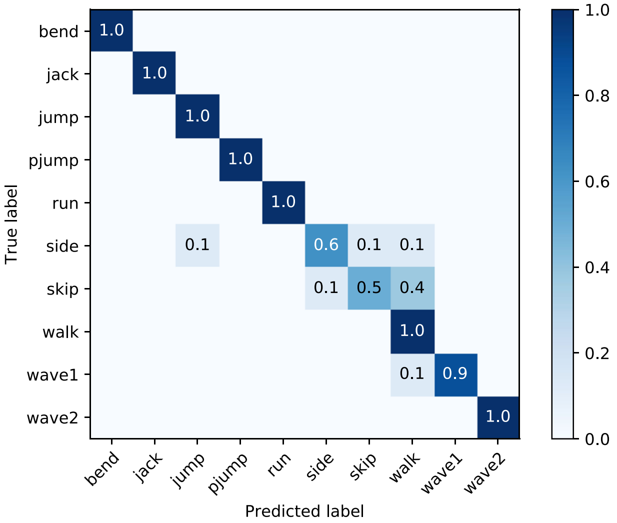

Fig. 6 Normalized confusion matrix for the classification of 10 actions of Weizmann Human Action dataset using one-shot learning

Table 3 Min. KL-Divergence values between training and testing actions using one-shot learning

		Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Test Actions	Bend	**2.14**	31.32	10.51	28.39	6.85	6.81	5.92	7.68	7.07	9.30
	Jack	3.77	**0.29**	3.64	1.35	2.90	3.09	4.20	4.15	4.00	3.02
	Jump	5.81	31.55	**1.94**	23.08	2.82	3.50	10.91	13.05	4.77	24.36
	Pjump	0.69	2.30	2.45	**0.69**	3.95	5.39	9.45	10.83	5.39	5.90
	Run	2.59	25.40	1.75	18.35	**0.50**	2.09	5.29	6.28	2.66	14.90
	Side	2.19	34.76	5.68	27.54	3.36	1.03	**0.86**	1.21	4.13	31.50
	Skip	2.27	23.00	11.77	13.55	5.29	3.31	**0.82**	0.85	8.10	25.57
	Walk	6.32	19.03	21.86	18.02	14.54	13.86	4.78	**4.15**	22.34	8.33
	Wave1	3.99	6.57	6.35	7.22	5.09	5.25	5.98	7.04	**1.62**	4.71
	Wave2	3.43	5.14	6.78	5.81	3.59	4.54	6.96	8.52	3.19	**1.74**

All testing actions are executed by “Daria” from the Weizmann Human Action dataset. Values highlighted in green correspond to correct classifications, whereas the one highlighted in red is an example of misclassification, in which the “side” action of Daria had higher similarity with the trained “skip” action than the trained “side” one

The lowest value of each row are highlighted in bold and colored in green if classified correctly, or red if misclassified

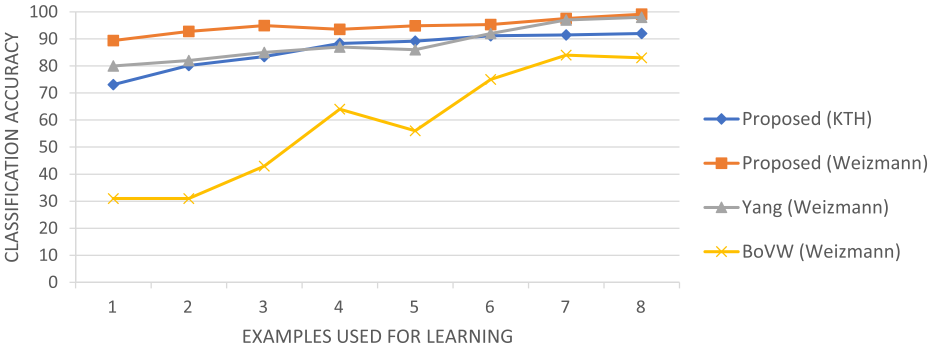

Fig. 7 Classification accuracy comparison between proposed method and others

5 Limitations and future work

This novel occurrence-based representation method that we have presented has been proven to be robust, even when only one training example is used. The use of KL-divergence values as a measure of similarity may also be combined with threshold values and used to discard outliers in human action datasets (e.g. Datasets which do not solely contain videos of human actions). For instance, a test video which has high KL-Divergence value with all trained actions would be labelled as an outlier. Regarding the time complexity of our method, since it is essentially based off an instance-based learning approach, it holds the same advantages and drawbacks as other lazy learning methods. The training phase is considerably efficient, but is coupled with a slow evaluation phase. Needless to say, the computation time highly depends on the nature of the application in which the method is being used and comes at the expense of some classification accuracy. For instance, a one-shot learning setting combined with minimal frame rescaling, high value of k (number of frames per sub-clip prior to application of KMeans) and low number of blocks in the KL-Divergence process, leads to lower computation time than a few-shot learning setting using opposite settings to maximize classification accuracy. Moreover, although the classification accuracies have been effective, there is room for improvement in the following sections:

Hyperparameters Different hyperparameters such as the GF-OF threshold t, the number of K-means components K and the number of Gaussian components n used per mixture, were set after conducting experiments to find their ideal values. Our next objectives include the automation of the adjustment of those hyperparameters, by designing both a feature extraction and a training model which can automatically adjust the hyperparameters according to the input data. For example, the training model would set the ideal number of Gaussian components n to represent a specific action and proceed through the “similarity measurement” process using that same number of components to find actions similar to the trained ones.

Unsupervised action recognition The training model that we have implemented was done in a supervised manner. Our next goal is to create a completely unsupervised human action recognition model which is capable of automatically finding action instances/repetitions within a same video and use one of those repetitions in the classification process. Additionally, we plan on utilizing the effectiveness of frameworks such as GAN and R-CNN to expand the flexibility of our work and enable its application in a wider range of datasets, including ones with multiple actors per video.

6 Conclusion

In this paper, an instance-based learning approach for human actions classification was proposed. The method employed Gunnar Farneback Optical Flow and K-means clustering to obtain accurate spatiotemporal features of an action, represented those features by a Gaussian mixture model, classified test videos using KL-divergence between two Gaussian mixtures and matched ones with the lowest divergence values. The conducted experiments involved validating an assumption made regarding the temporal perspective of each action instance, pinpointing the ideal number of Gaussian components to use per Gaussian mixture and running experiments using one-shot and *k*-shot learning. As displayed in the Section 4.2, the application of KL-Divergence as a similarity measure is demonstrated. Its computed values validate the usefulness of using such measure in our framework to not only achieve action classification, but to also give us a sense of how similar the actions in the dataset are. Similar actions exhibit low divergence values between each other, whereas dissimilar ones exhibit considerably higher values. The meaningful representation of human action instances, combined with the instance-based learning approach used, demonstrated that using as little as one training video per action yielded considerably high accuracies in comparison with state-of-the-art works. The flexibility of our work enables its application in other fields such as detection of outliers in datasets according to their KL-Divergence values (or similarity) with respect to the rest of the dataset. Additionally, various extensions could also be used on the proposed framework depending on the application, such as automating the hyperparameter tuning process used in the KMeans and GMM processes according to the input dataset to achieve higher classification accuracies while optimizing overall performance.

References

1. Avola D, Bernardi M, Foresti GL (2019) Fusing depth and colour information for human action recognition. Multimed Tools Appl 78:5919–5939
2. Blank M, Gorelick L, Shechtman E et al (2005) Actions as space-time shapes. In: Tenth IEEE International conference on computer vision (ICCV'05) Volume 1, vol 2, pp 1395–1402
3. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
4. Chao M, Xiaokang Y, Chongyang Z, Ming-Hsuan Y (2015) Long-term correlation tracking. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 5388–5396
5. Chen Y, Tao J, Liu L, Xiong J, Xia R, Xie J, Zhang Q, Yang K (2020) Research of improving semantic image segmentation based on a feature fusion mode. J Ambient Intell Humanized Comput 2020:1–13
6. Chen Y, Xu W, Zuo J, Yang K (2019) The fire recognition algorithm using dynamic feature fusion and IV-SVM classifier. Clust Comput 22(3):7665–7675
7. Christian S, Ivan L, Barbara C (2004) Recognizing human actions: A local SVM approach. In: Proceedings - International conference on pattern recognition, vol 3, pp 32–36
8. Darrell T, Pentland A (1993) Space-time gestures. In: Proceedings of IEEE conference on computer vision and pattern recognition. IEEE, pp 335–340
9. Farneback G (2000) Fast and accurate motion estimation using orientation tensors and parametric motion models. In: Proceedings 15th International conference on pattern recognition (ICPR-2000), vol 1, pp 135–139
10. Farneback G (2001) Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. In: Proceedings Eighth IEEE International conference on computer vision (ICCV 2001), pp 171–177
11. Fei Y, Li L, Binyong H et al (2019) Analysis and FPGA realization of a novel 5D hyperchaotic four-wing memristive system, active control synchronization, and secure communication application. Complexity 2019:1–18

12. Fei Y, Li L, Lin X et al (2019) A robust and fixed-time zeroing neural dynamics for computing time-variant nonlinear equation using a novel nonlinear activation function. Neurocomputing 350:108–116
13. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611
14. Goldberger J, Gordon S, Greenspan H (2003) An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. In: Proceedings Ninth IEEE International conference on computer vision, vol 1, pp 487–493
15. Hershey JR, Olsen PA (2007) Approximating the kullback leibler divergence between gaussian mixture models. In: 2007 IEEE International conference on acoustics, speech and signal processing - ICASSP '07, vol 4, pp IV-317-IV-320
16. Jianming Z, Chaoquan L, Jin W et al (2020) Training convolutional neural networks with multi-size images and triplet loss for remote sensing scene classification. Sensors 20(4):1188
17. Jianming Z, Zhipeng X, Juan S et al (2020) A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection. IEEE Access 8:29742–29754
18. João F H., Rui C, Pedro M, Jorge B (2014) High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 37(3):583–596
19. Kapsouras I, Nikolaidis N (2019) Action recognition by fusing depth video and skeletal data information. Multimed Tools Appl 78:1971–1998
20. Ke Y, Sukthankar R, Hebert M (2005) Efficient visual event detection using volumetric features. In: Tenth IEEE International conference on computer vision (ICCV'05) Volume 1, vol 1, pp 166–173
21. Kong Y, Fu Y (2018) Human Action Recognition and Prediction: A Survey. arXiv:1806.11230v2 [cs.CV]
22. Lin D, Weihong X, Yuantao C (2020) Density peaks clustering by zero-pointed samples of regional group borders. Comput Intell Neurosci 2020:8891778
23. Lu X, Ma C, Ni B, Yang X (2019) Adaptive region proposal with channel regularization for robust object tracking. IEEE Trans Circ Syst Video Technol 2019:1–1
24. Luoyu Z, Tao Z, Yumeng T, Hu H (2020) Fraction-order total variation image blind restoration based on self-similarity features. IEEE Access 8:30436–30444
25. Miller EG, Matsakis NE, Viola PA (2000) Learning from one example through shared densities on transforms, vol 1, pp 464–471
26. Najar F, Bourouis S, Bouguila N, Belghith S (2019) Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition. Multimed Tools Appl 78:18669–18691
27. Rodriguez M, Orrite C, Medrano C, Makris D (2016) Oneshot learning of human activity with an map adapted gmm and simplex-hmm. IEEE Trans Cybernet 47(7):1769–1780
28. Rodriguez M, Orrite C, Medrano C, Makris D (2017) Fast simplex-HMM for one-shot learning activity recognition. In: 2017 IEEE Conference on computer vision and pattern recognition workshops (CVPRW), pp 1259–1266
29. Seo HJ, Milanfar P (2011) Action recognition from one example. IEEE Trans Pattern Anal Mach Intell 33(5):867–882
30. Shechtman E, Gorelick L, Blank M, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
31. Wang H, Klaser A, Schmid C, Liu C (2011) Action recognition by dense trajectories. In: CVPR, pp 3169–3176
32. Xiankai L, Wenguan W, Chao M et al (2019) See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3623–3632
33. Xiankai L, Wenguan W, Jianbing S et al (2020) Learning video object segmentation from unlabeled videos. In: 2020 IEEE Conference on computer vision and pattern recognition (CVPR), pp 8960–8970
34. Yang Y, Saleemi I, Shah M. (2012) Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. IEEE Trans Pattern Anal Mach Intell 35(7):1635–1648
35. Yilmaz A, Shah M (2005) Actions sketch: a novel action representation. In: IEEE Computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 984–989
36. Yilmaz A, Shah. M (2005) Recognizing human actions in videos acquired by uncalibrated moving cameras. In: ICCV
37. Yuanjing L, Jiaohua Q, Xuyu X et al (2020) Coverless real-time image information hiding based on image block matching and dense convolutional network. J Real-Time Image Proc 17(1):125–135
38. Yuantao C, Jiajun T, Qian Z et al (2020) Saliency detection via the improved hierarchical principal component analysis method. Wirel Commun Mob Comput 2020

39. Yuantao C, Jie X, Weihong X, Jingwen Z (2019) A novel online incremental and decremental learning algorithm based on variable support vector machine. Clust Comput 22(3):7435–7445
40. Yuantao C, Jin W, Runlong X et al (2019) The visual object tracking algorithm research based on adaptive combination kernel. J Ambient Intell Humaniz Comput 10(12):4855–4867
41. Yuantao C, Jin W, Songjie L et al (2019) Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. Concurr Comput Pract Experience 2019:e5533
42. Yuantao C, Jin W, Xi C et al (2019) Single-image super-resolution algorithm based on structural self-similarity and deformation block features. IEEE Access 7:58791–58801
43. Yuantao C, Jin W, Xi C et al (2019) Image super-resolution algorithm based on dual-channel convolutional neural networks. Appl Sci 9(11):2316
44. Yuantao C, Linwu L, Jiajun T et al (2020) The improved image inpainting algorithm via encoder and similarity constraint. Vis Comput 2020:1–15
45. Zhang HB, Zhang YX, Zhong B et al (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19(1005):1–20

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.