2015 IEEE Winter Conference on Applications of Computer Vision

Action Recognition using Discriminative Structured Trajectory Groups

Indriyati Atmosukarto[1,2], Narendra Ahuja[3], Bernard Ghanem[4]
[1]Singapore Institute of Technology (SIT), Singapore
[2]Advanced Digital Sciences Center (ADSC), Singapore
[3]University of Illinois Urbana-Champaign (UIUC), USA
[4]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
indria@adsc.com.sg, n-ahuja@illinois.com, bernard.ghanem@kaust.edu.sa

Abstract

In this paper, we develop a novel framework for action recognition in videos. The framework is based on automatically learning the discriminative trajectory groups that are relevant to an action. Different from previous approaches, our method does not require complex computation for graph matching or complex latent models to localize the parts. We model a video as a structured bag of trajectory groups with latent class variables. We model action recognition problem in a weakly supervised setting and learn discriminative trajectory groups by employing multiple instance learning (MIL) based Support Vector Machine (SVM) using pre-computed kernels. The kernels depend on the spatio-temporal relationship between the extracted trajectory groups and their associated features. We demonstrate both quantitatively and qualitatively that the classification performance of our proposed method is superior to baselines and several state-of-the-art approaches on three challenging standard benchmark datasets.

1. Introduction

Human action recognition is a very challenging computer vision problem. Key challenges in action recognition include subject differences, background clutter, occlusion, and appearance variation. Actions belonging to the same class performed by two different persons frequently look significantly different, while actions belonging to different classes may look very similar. A popular current approach to action recognition in videos is to represent the videos using a bag-of-feature (BOF) derived from the spatio-temporal volume and to classify using non-linear Support Vector Machines (SVM), while ignoring the spatial and temporal structure of the features.

In this paper, we propose to represent a video in terms of groups of trajectories characteristically associated with the action, through class variables (Figure 1). The goal is to determine the trajectory groups that are most discriminative of an entire action or a specific, signature part of the action, and to use these groups to learn a model for each action. We use trajectories because we believe that trajectories, by capturing both motion and location information, are more representative as well as discriminative compared to interest point based descriptors. Each trajectory group is cast as an instance in a multiple instance learning (MIL) framework. The selected trajectory groups need to be representative to capture the high variance within the same action class and discriminative so that they can be used to differentiate between the different classes. We model the spatio-temporal relationships between trajectory groups using graphs. The learned discriminative trajectory groups correspond to cliques of co-existing trajectory groups. This framework can also be applied to other video descriptors, including interest point based descriptors.

Our work draws inspiration from part-based model in object recognition [7] and part-based model in action recognition [18, 14, 9, 10]. Unlike previous approaches, our approach avoids the need for complex matching, complex latent models and expensive iterative optimization algorithms to learn parameters. Our approach also avoids time consuming matching using sliding image or volume windows.

The contributions of this work are as follows: i) We propose a framework based on trajectory groups to solve the action recognition problem; ii) We adopt a weakly supervised setting, where videos are represented as bags of trajectory group instances with latent class variables; iii) We design a graph kernel from spatio-temporal structure information, to enable multiple instance learning. We compare our results to those obtained by state-of-the-art, parts-based as well as non-parts-based approaches, and show that our approach performs better on standard benchmark datasets.

We review related work in Section 2. We proceed to describe our video representation in Section 3. We describe the evaluation of our approach and several state-of-the art approaches in Section 4. We conclude and discuss future work in Section 5.

978-1-4799-6683-7/15 $31.00 © 2015 IEEE
DOI 10.1109/WACV.2015.124

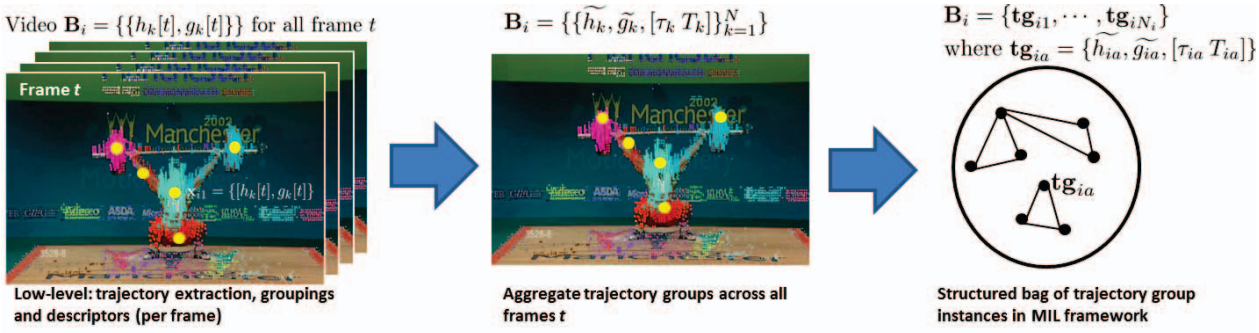

Figure 1. Our proposed approach represents a video as a bag of (instances of) trajectory groups as nodes, along with an undirected graph whose connections capture the spatiotemporal relationships among the trajectory groups. Instances that co-exist in the video (occur in the same time interval) are connected, while instances that do not co-exist are not connected.

2. Related Work

Spatio-temporal vs Trajectory-based approach. Previous approach for action recognition are based on bag-of-features (BOF) by quantizing spatio-temporal volume features [20, 11]. The BOF-approach detects spatio-temporal interest points and extract descriptors such as histogram of oriented gradient (HoG), histogram of optical flow (HoF), HoF3D, and histogram of oriented edges of motion boundaries (HoMB). These local low level features are unfortunately not discriminative enough. Other works have focused on trajectory-based approach, instead of extracting spatio-temporal interest points, they use trajectories to describe the video [19]. Yi et al. [25] extracted salient trajectories by considering both their appearance and motion saliency. Similarly, Atmosukarto et al. [4] extracted trajectories and used Fisher Kernel to create video representation.

Mid-level Video Representation. Riding the success from object recognition field, action recognition field has now started to investigate part-based model. Gaidon et al. [1] represented a video as a hierarchy of trajectories (tracklets). The hierarchy is first computed using a divisive clustering algorithm and then represented as a BOF-tree structure where each node is modeled by a bag-of-feature over Motion Boundary Histograms (MBH) descriptors. Raptis et al. [14] considered both recognition and localization when modeling their approach. They first extracted clusters of long term trajectories and learned a latent model over a fixed number of part. They used graphical models to represent the structure and relationship between the parts. Unfortunately, their approach had cubic time complexity in terms of number of trajectories as a result only a small subset of the clusters (three) are automatically selected to represent the video. Tian et al. [18] generalized deformable part models from 2D images to 3D spatiotemporal volumes, where the most discriminative 3D subvolumes are automatically selected. Jain et al. [9] proposed to use discriminative spatio-temporal patches to represent videos, where the patches are learned in an unsupervised manner using exemplar-SVM to cluster the data. Wang et al. [21] proposed a mid-level representation (part-based) for video called motionlet activation vector. Motionlets are spatio-temporal (3D) regions (or parts) with stable features in both in motion and appearance space. These parts are learned in an unsupervised manner.

Multiple Instance Learning (MIL). Most approaches to MIL assume that instances of a bag are independent. This assumption is inappropriate for many applications where structural and spatial dependencies exists, for example image processing and video analysis. Few recent work have started to introduce spatial information to their MIL formulation. Warrell et al. [22] incorporate MIL constraints into Conditional Random Field (CRF) models to construct structured bag models that capture the spatial dependencies among instances in the bag. Their approach was tested on two vision tasks: image labeling and image segmentation. Most similar to our work is the work of Zhou et al. [27]. Their approach exploited the relations among instances by implicitly constructing graphs through deriving affinity matrices and proposing an efficient graph kernel that considers clique information. However, their method binarizes the feature distances of instances where the relational information between instance features is reduced to sum of binarized distances. Herman et al. [8] proposed a multi-instance kernel approach, MIRKernel, that takes into account the relational information of instances when computing similarities betwen bags. Their approach mainly used statistics derived from feature distances to form relational information. Cuingnet et al.[6] introduce an SVM framework to spatially regularize SVM for brain image analysis. They show that Laplacian regularization provides a flexible framework that can integrate spatial and anatomical constraints and show its application to classification of MR images.

Action Recognition and MIL. Ali et al. [2] proposed a set of kinematic features that are derived from optical flow for human action recognition in videos. A video is represented as a bag of kinematic modes. Closely related to our work is the work of Sapienza et al. [16] where discriminative action subvolumes are learned in a weakly supervised

setting. The learned models are used to classify and localize actions. Each video is decomposed into subvolumes and as a result a video is represented as a bag of histograms. Instead of using subvolumes representation, we use trajectories and extract trajectory groups [14] from the video and aim to learn the discriminative trajectory groups to represent the video. Most importantly, our representation maintains the structural spatio-temporal information in each bag, while [16] treated each instances in a bag independently.

3. Video Representation

Our proposed solution aims to recognize human actions in video clips. Our framework is general and can be applied to any video descriptor including interest point based descriptors. In this paper, we adopt trajectory-based representation as we believe trajectories contain more discriminative information on motion and action in videos.

3.1. Low-Level Feature Extraction

Trajectories are efficient in capturing object motions and actions in videos. We use trajectory extraction, groupings and descriptors proposed by Raptis et al. [14]. In their approach, dense trajectories [5] are clustered using an efficient greedy agglomerative hierarchical clustering algorithm [17]. While clustering, a spatial threshold is applied to each pair of trajectory to ensure spatial compactness of the groups; pairs of trajectory whose distances are above the threshold are not grouped into the same trajectory group [14]. At every frame, every trajectory is assigned a membership label to a particular trajectory group. Every trajectory in a group is spatially quantized to a regular grid and its descriptor is quantified with a pre-learned codebook label. Each trajectory group k is finally described by the concatenation of the histogram of oriented gradients (HoG), histogram of optical flows (HoF), and histogram of oriented edges of motion boundaries (HoMB) for every frame t in which the trajectory members exist; we denote this descriptor as $h_k[t]$. A mean group trajectory is computed by calculating the average position $g_k[t]$ of all trajectories in group k at each time t. The time interval in which the group exists is marked and denoted by $[\tau_k\ T_k]$, to indicate the starting frame and ending frame for each trajectory group k.

3.2. Discriminative Trajectory Groups

We represent a video in terms of clusters of trajectory groups, obtained as described in Sect. 3.1. Our goal is to learn discriminative trajectory groups that are most relevant to a specific action and to use them to learn a model for each action class. We aggregate the low-level trajectory group descriptors $h_k[t]$ and $g_k[t]$ in terms of their average values across the time interval $[\tau_k\ T_k]$ in which group k exists. The group averages are denoted as $\widetilde{h_k}$ and $\widetilde{g_k}$. Consequently, a video $\mathbf{B}$ is described as the collection of all trajectory groups k, that is $\mathbf{B} = \{\{\widetilde{h_k}, \widetilde{g_k}, [\tau_k\ T_k]\}_{k=1}^{N}\}$ where N is the total number of trajectory groups that exist in the video.

We cast our learning task in a multiple instance learning (MIL) framework where every video is a bag and the trajectory groups extracted from the video are instances within the bag. Instances in a bag are trajectory groups that exist at any point in time in the video. We implicitly construct an undirected graph for every bag where the nodes of the graph are the trajectory group instances and the edges are the spatio-temporal co-existing properties of the instances. The resulting graph may not be fully connected as not all instances co-exist throughout the video. A clique in the graph indicates instances (trajectory groups) that co-exist in time (Fig. 1). We aim to model the interaction among instances in a bag and propose to define graph kernels that consider the structural information between instances inside a bag.

A video (bag) $\mathbf{B}_i = \{\mathbf{tg}_{i1}, \cdots, \mathbf{tg}_{iN_i}\}$ is a collection of trajectory groups (instances), where each group is described by its features, position and time interval, that is $\mathbf{tg}_{ia} = \{\widetilde{h_{ia}}, \widetilde{g_{ia}}, [\tau_{ia}\ T_{ia}]\}$. Given a training set $D = \{(\mathbf{B}_i, Y_i), \ldots, (\mathbf{B}_m, Y_m), i = 1, 2, \ldots, m\}$ where $Y_i \in \{-1, 1\}$ is the label of a video bag $\mathbf{B}_i$, the goal is to learn the label for each trajectory group (instance in a bag) and an action model to represent each action class.

The structural information of an instance captures the context in which the instance occurs in the bag. We define the structural information of an instance relative to other instances from the same bag by considering two different similarity measures between instances: (1) feature similarity and (2) relative position. Note that we did not take compactness of the trajectory groups as a measure of goodness and factor to consider when calculating instance similarity as this factor was already taken into account when forming the trajectory groups as mentioned in Sect. 3.1.

We design multi-instance kernel for the different similarity properties: kernel k_1 to measure the feature similarity between instances and kernel k_2 to measure pairwise spatial distance between temporally co-existening instances. We define kernel k_1 as

$$k_1(\mathbf{tg}_{ia}, \mathbf{tg}_{jb}) = \exp(-\gamma_1 ||\widetilde{h_{ia}} - \widetilde{h_{jb}}||^2) \qquad (1)$$

where $\mathbf{tg}_{ia}$ is the trajectory group a for video i, while $\mathbf{tg}_{jb}$ is the trajectory group b for video j. $\widetilde{h_{ia}}$ and $\widetilde{h_{jb}}$ are the feature descriptors for trajectory group $\mathbf{tg}_{ia}$ and $\mathbf{tg}_{jb}$.

We define the pairwise spatial distance measure for two co-existing instances as

$$d(\mathbf{tg}_{ia}) = \{||\widetilde{g_{ia}} - \widetilde{g_{ib}}|| | \forall_{\mathbf{tg}_{ib} \in \mathbf{B}_i}, b \neq a \text{ and } [\tau_{ia} T_{ia}] \cap [\tau_{ib} T_{ib}]\} \qquad (2)$$

where $\widetilde{g_{ia}}$ is the mean group position for trajectory group a for video i, similarly for $\widetilde{g_{ib}}$ for trajectory group b in video i. τ_{ia} is the starting frame for trajectory group a in video i, while T_{ia} is the ending frame for trajectory group a in video i. Similarly for τ_{ib} and T_{ib}. We set the distance between an instance to itself to be 1. We normalize the distances by the maximum pairwise spatial distance between bag instances.

Kernel k_2 is then defined as

$$k_2(d(\mathbf{tg}_{ia}), d(\mathbf{tg}_{jb})) = \exp(-\gamma_2 ||d(\mathbf{tg}_{ia}) - d(\mathbf{tg}_{jb})||^2) \quad (3)$$

We formulate the final kernel for measuring the similarity between video $\mathbf{B}_i$ and video $\mathbf{B}_j$ as follows:

$$K_{TG}(\mathbf{B}_i, \mathbf{B}_j) = \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} \{k_1(\mathbf{tg}_{ia}, \mathbf{tg}_{jb}) k_2(d(\mathbf{tg}_{ia}), d(\mathbf{tg}_{jb}))\} \quad (4)$$

We normalize the final kernel to

$$K_{TG}(\mathbf{B}_i, \mathbf{B}_j) = \frac{K_{TG}(\mathbf{B}_i, \mathbf{B}_j)}{\sqrt{K_{TG}(\mathbf{B}_i, \mathbf{B}_i) K_{TG}(\mathbf{B}_j, \mathbf{B}_j)}} \quad (5)$$

3.3. Action Recognition

Given a training dataset, we obtain an action model for each human action class. We first extract the trajectory groups from each video in the training set (Sect. 3.1). We then compute the pairwise kernel matrix for all videos in the training set and use the pre-computed kernels to learn a one-vs-all SVM model for each action class as discussed in Sect. 3.2. Given a new test video, we follow the same framework, extract the trajectory groups and compute the kernel matrix for the test data to evaluate the test video against each of the action models. Final label is chosen as the action model that gives the highest probability score.

4. Experiments

In this section, we present experimental results to validate the effectiveness of our approach. We measure performance on three standard action recognition benchmark datasets, in quantitative as well as qualitative terms.

4.1. Datasets

The **Hollywood (HOHA) dataset** [11] consists of 430 movie clips from 32 movies. The clips are categorized into 8 classes: Kiss, StandUp, SitDown, AnswerPhone, HugPerson, Handshake, GetOutCar, SitUp. This dataset is challenging due to the uncontrolled settings of the video clips resulting in change of illumination, and heavy background.

The **UCF Sports dataset** [15] consists of almost 200 video clips. The clips consist of sports actions from broadcast television, with varying scenes and viewpoints, making the dataset very challenging. The clips are categorized into 10 classes: dive, golf swing, kick, lift, horse riding, run, skate, swing-on-bench, swing-on-bar, and walk.

The **TV Human Interaction (TVHI)** dataset [13] consists of 300 video clips containing one of four interactions: handshake, highfive, hug, and kiss. In addition, 100 videos containing no interactions were added to represent the negative class. The dataset has high intra-class variation to capture realistic settings in human interactions.

4.2. Experiment Setup

We used the implementation and default parameters provided in [14] to extract the trajectory groups from all the videos in the dataset.

The **HOHA dataset** consists of 219 training and 217 testing videos. Using the method described in Sect. 3.1, on average we extracted 148 trajectory groups per video from the training dataset and 213 trajectory groups per video from the test dataset. The average number of movie frames in each of the training and test sets is 163 frames. Due to the large number of trajectory groups per video, we randomly selected N (=50) trajectory groups per video. Trajectory groups that existed for a longer time interval in the video had a higher probability of being sampled, the assumption being that these trajectory groups are more relevant. The probability that a trajectory group $\mathbf{tg}_{ia}$ is sampled is $(T_{ia} - \tau_{ia})/T_i / \sum_{a=1}^{n_i} (T_{ia} - \tau_{ia})$ where T_i is the duration of the whole video. Similar to Raptis et al. [14] and their experimental setup for the HOHA dataset, we use the bounding box annotation to aid the training phase by selecting trajectory groups that are relevant to the action. In the case of unbalanced dataset such as HOHA, where the number of positive classes is much smaller than the number of negative classes, we replicate the positive samples to be equal to the number of negative samples. We noticed that this practice increases the classification accuracy by at least 2% [26].

We used the same train-test dataset split for the **UCF Sports dataset** as in the work of Lan et al. [10] and Tian et al. [18]. Instead of a leave-one-out methodology, the dataset is split by taking one third of the videos from each action class to form a test set, and the rest is used for training. This split results in 103 training videos and 47 test videos. On average, we extracted 358 trajectory groups per video from the training dataset and 330 trajectory groups per video from the test dataset. The average number of movie frames in both the training and test sets is 63 frames. We sampled $N = 50$ trajectory groups to represent each video.

The **TVHI dataset** is split into 150 training and 150 test videos. Using the method described in Sect. 3.1, on average we extracted 216 trajectory groups from the training videos and 233 trajectory groups from the test videos. We sampled $N = 50$ trajectory groups to represent each video. The average number of frames in the videos is 90. For all datasets, the feature descriptors were normalized to $[0, 1]$.

We compare our performance results with those obtained by both part-based and non-part-based state-of-the-art action recognition methods [27, 14, 10, 20, 21]. In addition, we also define two baseline methods for comparison.

Baseline A In our first baseline experiment, we compare our multi-instance structural kernel approach to mi-SVM [3] which is a multi-instance kernel method that assumes that there exists no relation between instances in the bag. We used an off-the-shelf mi-SVM implementation (MILL) [24]. This baseline is also equivalent to precomputing the kernel but ignoring the spatio-temporal relationships between instances in the bag.

Baseline B In our second baseline experiment, we modify our spatio-temporal kernel calculation to include coexistence of a pair of trajectory groups but not the spatial distance. That is $d(\mathbf{x}_{ia})$ in Eq. 2 is set to 1 for two coexisting instances in the bag, and set to 0 otherwise. This is similar to the graph kernel representation from Zhou et al. [27]

In all our experiments, the kernel parameters γ_1 and γ_2 and SVM parameter C were selected based on five-fold cross-validation.

4.3. Experimental Results

Action Recognition results. For the **HOHA** dataset, we adopt experimental settings similar to those in [14] where we compute the average precision (AP) to evaluate the performance of the different methods. Table 1 shows the performance comparison between our method and our two different baselines on the HOHA dataset. Our approach performs better than both baselines. This shows that incorporating the spatio-temporal structure into the kernels improves classification accuracy. The performance results of our method on the HOHA dataset is lower than those of Raptis et al. [14] where Raptis et al. [14] obtained mean AP of 0.33 by using a BoW representation and SVM with RBF χ_2 kernel. This can be explained by the high intra-class variance in the action in this dataset. Wu et al. [23] applied low rank optimization to decompose trajectories and obtained an average precision of 0.476, however in their experiments they used a 'clean' training set where videos with shot changes were removed from the training set.

Table 1. Comparison of classification performance on the Hollywood dataset (mAP = mean Average Precision; avgAcc = average Accuracy). Note that random avgAcc classification for this dataset is at 0.125.

Method	Evaluation	Result
Baseline A	avgAcc	0.14
Baseline A	mAP	0.207
Baseline B	avgAcc	0.31
Baseline B	mAP	0.212
Our method	avgAcc	0.34
Our method	mAP	**0.252**

For the **UCF-Sports** dataset, we train a model for each of the ten action classes and use a one-versus-all classification evaluation to obtain the average classification accuracy. Table 2 summarizes the classification performance of our method compared to our two different baselines and several state-of-the-art approaches. Our recognition results on this dataset are competitive. Our method performs better than all the baselines and state-of-the-art methods. We can see that baseline A performs better than the BOW method by 2%, while baseline B performs better than the method proposed by Lan et al. [10]. By adding spatio-temporal structural information when constructing graphs for our MIL framework, we see a 13% improvement over baseline A and 9% over baseline B. Figure 2 shows the per-class classification accuracy on the UCF Sports dataset. Comparing the results of baseline A to those of baseline B and our method, the results verify our expectation that addition of spatio-temporal structural information improves the classification accuracy for most of the classes. We also compare our method to [20, 19] and [21]. In [20], Wang et al. proposed a combination of dense HOG3D feature descriptor and bag-of-feature SVM approach and obtained average accuracy of 85.6%. Subsequently, in [21] they added spatio-temporal pyramid to their representation and achieved 89.1% average accuracy. The method in [19] used dense trajectories and bag-of-feature SVM approach and obtained an average accuracy of 88.2%. However, all three works extended the original training dataset to include a horizontally flipped version of each sequence in the dataset and used a leave-one-actor setup which is known to have higher accuracy than random cross validation. The confusion matrix for the different classes in the UCF Sports dataset using our approach is shown in Figure 3.

Table 2. Performance comparison for average classification accuracy on the UCF-Sports dataset.

Method	Accuracy (%)
BoW [14]	67.4
Lan et al. [10]	73.1
Tian et al. [18]	75.2
Raptis et al. [14]	79.4
Baseline A	69.6
Baseline B	73.9
Our method	**82.6**

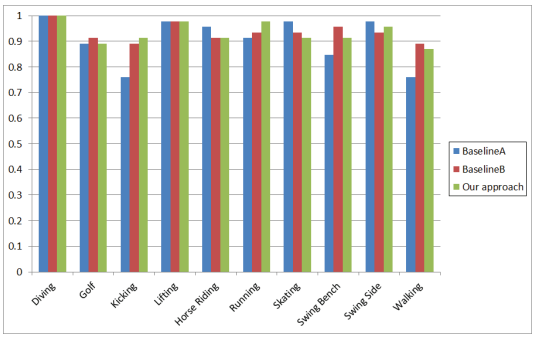

Figure 2. Per-class classification accuracy on UCF Sports dataset. Our method achieves an average accuracy of 84.78% ($N = 100$), an increase of 8% in performance compared to state-of-the-art methods.

Results show that our approach performs better than existing non-parts-based model [10], parts-based model approaches [14, 18] and graph kernel approaches [27]. We can conclude that our video representation based on discriminative structured trajectory groups is effective in performing

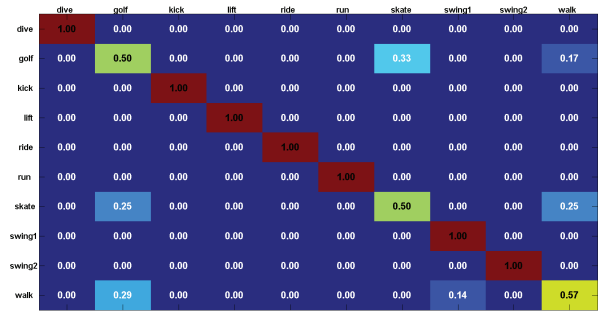

Figure 3. Confusion matrix using our proposed approach on the UCF Sports dataset.

action recognition even for challenging dataset settings such as HOHA. The results also demonstrate the usefulness of relational information of instances in MIL, and specifically, of spatio-temporal structural information, compared to approaches that are based on local features such as HOG/HOF.

Visualization of trajectory groups. We present several qualitative results on the datasets. We first visualize the trajectory groups that were sampled based on their durations in the video. Weighted sampling was performed to avoid trajectory groups with short duration in the video. Short trajectories might play an important role for fine-grained action recognition where the goal is to distinguish between similar actions. However for our purposes we are distinguishing very different actions and therefore subtle differences (eg. short trajectories) can be treated as irrelevant. Figure 4 shows all trajectory groups that exist in sample frames from test videos in the HOHA dataset (the mean trajectory groups are marked with plus signs). The selected trajectory groups are marked by yellow plus signs, while the rest of the trajectory groups are marked with red plus signs. We can see that most of the selected sampled trajectory groups (in yellow) are relevant to the action.

When constructing the graph for our Multiple Instance Learning (MIL) framework, trajectory groups that co-exist in the video will ultimately form a clique in the graph. Here, we visualize members of the different cliques in the graph. From our experiments, we observe that trajectory groups that form the largest clique in the graph are the trajectory groups that are most relevant to the action, an observation which is equivalent to action localization.

The UCF Sports dataset is not very appropriate to evaluate action localization as the relevant action occurs throughout the whole video clip. As such we show our qualitative results on the HOHA dataset. Figure 5 (top row) shows two sample frames highlighting the trajectory groups that form two different cliques for a test video labeled 'GetOutCar'. For this particular video, four cliques were formed when constructing the graph. The first clique (top row, left) consists of 38 trajectory group members (marked by yellow plus signs on the frame), while the second clique (top row, right) consists of only 6 members. The remaining cliques have only 5 and 1 members respectively. The largest clique corresponds to the discriminative structured trajectory groups. Figure 5 (bottom row) shows montages of the common frames extracted from members of the cliques in Figure 5 respectively. The montage is created by selecting the most common frames among all the trajectory group members in the clique. From the montages in Figure 5, we can see that the largest clique (bottom row, left) corresponds to trajectory groups that are relevant to the action ('GetOutCar'). The second montage (bottom row, right) shows that the second largest clique corresponds to trajectory groups that are less relevant to the action (the actor has already exited the car). Figure 6 shows more examples of the structured discriminative trajectory groups and their respective montages. The results show that the structured discriminative trajectory groups that form the largest clique during our graph construction approach are more relevant to the action and can be used to localize the action.

AnswerPhone Handshake

Figure 4. Weighted sampling of trajectory groups in a video. The trajectory groups were sampled based on their durations in the video. The selected sampled trajectory groups (marked by yellow plus sign), which are sampled due to their longer durations appear to be more relevant to the action, confirming our argument that short trajectory groups tend to correspond to noise or background motion in the video.

Varying number of sampled trajectory groups. We performed additional experiments by varying the size of the graph representation of a video bag. To investigate the impact of the number of trajectory groups sampled to represent a video, we varied the sample size ranging from 50 trajectory groups per video to all extracted trajectory groups in the video. The average number of sampled trajectory groups in the UCF Sports dataset is 350. Figure 7 shows the effect of the number of sampled trajectory groups to the average accuracy on the UCF Sports dataset. The average accuracy for the UCF Sports dataset when the number of sampled trajectory groups N is set to 50 was 82.61%. The average accuracy increased to 84.78% when N was increased to 100. However, increasing the number of sampled trajec-

Figure 5. (Top row) Structured discriminated trajectory group instance form the largest clique in the graph within a video bag in our MIL framework (left). The second largest clique (right) occurs in frames that are not as relevant to the action. (Bottom row) Montages showing the common intersecting frames from members in the clique. The first montage (left) shows frames that are more relevant to the action ('GetOutCar'). In the second largest clique (right) the frames are not as relevant to the action (The actor has already exited the car in this example).

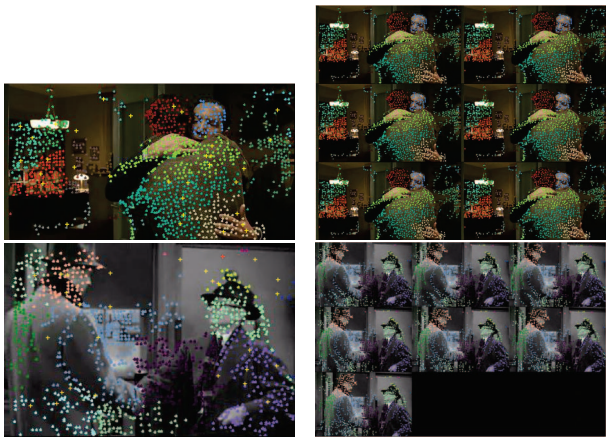

Figure 6. More examples of the structured discriminative trajectory group (marked by yellow plus sign) and the montage showing the common intersecting frames. The results show that the trajectory groups that form the largest clique in the graph are most relevant to the action ('Hug' and 'HandShake' in this example).

tory groups to 150 or higher did not lead to an improvement in the average classification accuracy. This might be due to the fact that by increasing the number of trajectory groups too much, we are including irrelevant trajectory groups that correspond to noise or background motion.

The time spent on computing the kernels increases as the number of sampled trajectory groups increases. Our experiments were performed on an Intel(R)Xeon(R) CPU W3530@2.8GHz. The time for computing the test kernel for the whole test dataset is 1 minute per action class when the number of sampled trajectory group, N, is set to 50.

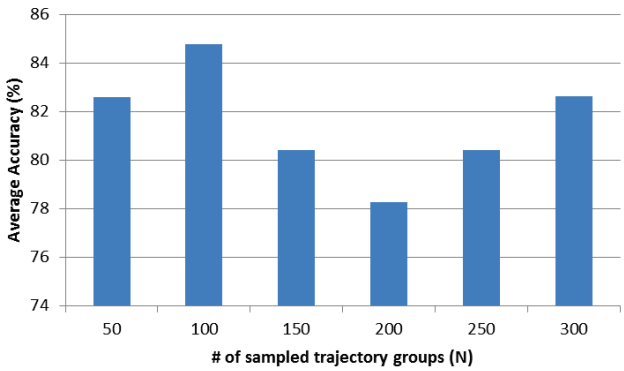

Figure 7. Effect of varying number of sampled trajectory groups (N) on action recognition average accuracy for UCF Sports dataset.

The computation time increased to a total of 3 minutes for $N = 100$ and further increased to 7 minutes for $N = 150$. Note that this computation time is still much cheaper compared to complex graph matching or sliding volume approaches. The method in [14] has to restrict the number of parts that make up an action due to the complexity of matching between graphs which can be NP-complete. Restricting the number of parts to such a small number (three parts in their experiments) is very constraining given most actions in semantic settings would have more than three parts.

We analyzed the effect of duration of co-existence of the trajectory groups when calculating the graph kernel. The results showed not much improvement in the classification accuracy. When using the length of overlap between trajectory groups during graph kernel computation, we achieved similar classification accuracy of 82.61% with $N = 50$ using cross-validation for parameters C, γ_1, and γ_2.

Recognition of human interactions. Most of the action classes in common action recognition benchmark datasets are autonomous, that is the action is performed by only one actor and is independent of other actors in the video. We performed additional experiments to investigate the performance of our method in recognizing interactions between two people in a realistic scenario. Only three of the 8 classes (Kiss, HugPerson and Handshake) in the HOHA dataset represent some form of interactions. We performed additional experiments on the **TVHI dataset** which was compiled to represent realistic interactions. Table 3 summarizes the mean average precision (mAP) performance on the dataset. The experiments show promising results and shows that our approach is able to capture interactions between actors. The results highlight the strengths of part-based model vs. non-part-based model when used to recognize finer actions such as human interactions. Our method performs slightly worse than the recent work of Patron-Perez et al. [12]. However, their approach is person-centric and requires dedicated tracking of upper bodies and heads in the

video. Gaidon et al. [1] achieve 55.6% classification accuracy on the TVHI dataset, however their method require spectral divisive clustering which is much more computationally expensive than our method.

Table 3. Performance comparison for mean Average Precision (mAP) on the TVHI dataset.

Method	mAP (%)
Patron-Perez et al. [13]	32.8
Laptev et al. [11]	36.9
Patron-Perez et al. [12]	42.44
Our method	41.47

5. Conclusion

In this paper we have presented a method for human action recognition using discriminative structured trajectory groups. This work is inspired by part-based models approach. We represent a video as a set of trajectory groups described by their flow and motion features. We cast our method into an MIL framework where each video is a bag and each trajectory group is an instance in the video bag. We consider the spatio-temporal relationship between each trajectory group in the video by looking at the spatial distance and temporal coexistence when building a graph to represent the video bag. We demonstrate that this representation can be used to obtain action recognition results equivalent to or better than state-of-the-art methods on challenging action recognition benchmark datasets. Our approach avoids the need for complex matching, complex latent models and expensive iterative optimization algorithms to learn parameters. We also avoid the use of sliding image or volume window approaches which are time consuming. Future work include modeling human interactions more concretely.

6. Acknowledgement

This study is supported by the research grant for the Human Sixth Sense Programme at ADSC from A*STAR.

References

[1] C. S. A. Gaidon, Z. Harchaoui. Recognizing activities with cluster-trees of tracklets. In *BMVC*, 2012. 2, 8

[2] S. Ali and M. Shah. Human action recognition in videos using kinematic features and MIL. *PAMI*, 32:288 – 303, 2010. 2

[3] S. Andrews, I. Tsochantaridis, and T. Hoffman. Support vector machines for MIL. In *NIPS*, 2002. 4

[4] I. Atmosukarto, B. Ghanem, and N. Ahuja. Trajectory-based fisher kernel representation for action recognition in videos. In *ICPR*, 2012. 2

[5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 3

[6] R. Cuingnet, J. A. Glaunes, M. Chupin, H. Benali, O. Colliot, and T. A. D. N. Initiative. Spatial and anatomical regularization of svm: A general framework for neuroimaging data. *PAMI*, 35(3), 2013. 2

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 1

[8] G. Herman, G. Ye, Y. Wang, J. Xu, and B. Zhang. Multi-instance learning with relational information of instances. In *WACV*, 2009. 2

[9] A. Jain, A. Gupta, M. Rodriguez, and L. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2010. 1, 2

[10] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 1, 4, 5

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2, 4, 8

[12] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *PAMI*, 2012. 7, 8

[13] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, 2010. 4, 8

[14] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 1, 2, 3, 4, 5, 7

[15] M. D. Rodriguez, J. Ahmed, and M. Shah. Actionmach = a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 4

[16] M. Sapienza, F. Cuzzolin, and P. H.S.Torr. Learning discriminative space-time actions from weakly labelled videos. In *BMVC*, 2012. 2, 3

[17] S. Tabatabaei, M. Coates, and M. Rabbat. Ganc: Greedy agglomerative normalized cut. *arXiv:1105.0974*, 2011. 3

[18] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 2, 4, 5

[19] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2, 5

[20] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2, 4, 5

[21] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *CVPR*, 2013. 2, 4, 5

[22] J. Warrell and P. H. Torr. Multiple-instance learning with structured bag models. In *EMMCVPR*, 2011. 2

[23] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, 2011. 5

[24] J. Yang. Mill: A multiple instance learning library. 4

[25] Y. Yi and Y. Lin. Human action recognition with salient trajectories. *Signal Processing*, 93:2932–2941, 2013. 2

[26] Z.-Q. Zeng and J. Gao. Improving svm classification with imbalance data set. In *NIPS*, 2009. 4

[27] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. MIL by treating instances as non-iid samples. In *ICML*, 2009. 2, 4, 5