CollectiveSports: A Multi-task dataset for collective activity recognition

Cemil Zalluhoglu*, Nazli Ikizler-Cinbis

Department of Computer Engineering, Hacettepe University, Ankara, Turkey

Abstract

Collective activity recognition is an important subtask of human action recognition, where the existing datasets are mostly limited. In this paper, we look into this issue and introduce the "Collective Sports (C-Sports)" dataset, which is a novel benchmark dataset for multi-task recognition of both collective activity and sports categories. Various state-of-the-art techniques are evaluated on this dataset, together with multi-task variants which demonstrate increased performance. From the experimental results, we can say that while sports categories of the videos are inferred accurately, there is still room for improvement for collective activity recognition, especially regarding the generalization ability beyond previously unseen sports categories. In order to evaluate this ability, we introduce a novel evaluation protocol called *unseen sports*, where the training and test are carried out on disjoint sets of sports categories. The relatively lower recognition performances in this evaluation protocol indicate that the recognition models tend to be influenced by the surrounding context, rather than focusing on the essence of the collective activities. We believe that C-Sports dataset will stir further interest in this research direction.

Keywords: Collective Activity Recognition, Action Recognition, Convolutional Neural Networks, Multi-task learning, LSTM

*Corresponding author

Email addresses: cemil@cs.hacettepe.edu.tr (Cemil Zalluhoglu), nazli@cs.hacettepe.edu.tr (Nazli Ikizler-Cinbis)

1. Introduction

Recognition of collective activities, which is defined as the collective behaviour of multiple people in a scene, is a recent topic of interest in computer vision community. The task has various application domains, ranging from patient monitoring to collective sports analysis, to large scale surveillance and beyond. Despite the wide range of applicability, collective activity recognition is a relatively less-studied topic, compared to human action recognition. Moreover, the existing datasets are mostly limited, in the sense that they are not diverse enough to support the training of complex and representative models.

In this paper, we address this shortcoming for collective activity recognition and present a novel collective activity dataset, called "Collective Sports (C-Sports for short)", which includes various collective activities occurring in multiple sports videos. This dataset has an interesting property of being multi-task in nature. Specifically, we collect several common collective activities, such as *gathering*, *dismissal*, *attack*, etc., that take place in various sports matches. Different from the existing related datasets such as Volleyball dataset [1], our dataset consists of videos ranging from a diverse set of sports, from basketball to dodgeball, to ice hockey or waterpolo. This diversity makes the dataset more interesting and compelling at the same time, from the recognition point of view.

A brief comparison of existing collective activity recognition datasets is presented in Table 1. Firstly, as it can be seen, the number of collective activity datasets is quite limited. Collective Activity Dataset (CAD), introduced in [2], consists of only 44 sequences of 5 collective activities. The people in this dataset is mostly seen orthogonally at relatively near distance. New Collective Activity Dataset (nCAD) [3], which is composed of 32 video clips with 6 collective activities, consists of artificially posed sequences. The more recent Volleyball dataset[1], on the other hand, consists of collective activities that occur only in Volleyball sport, therefore, the domain is limited to Volleyball activities only. This limitation is likely to hinder the generalization ability of the methods tuned for this dataset to other domains directly.

Table 1: Comparison of the collective activity video datasets in the literature.

Dataset	**#CollActs**	**#Category**	**#Videos**	**#Frames**	**View point**	**Camera Movement**
CAD[2]	5	N/A	44	25756	near	stationary
nCAD[3]	6	N/A	32	19873	near	stationary
Volleyball[1]	6	1	1636	67076	far	stationary
C-Sports	5	11	2187	167935	near/far	non-stationary

In order to address such limitations, C-Sports dataset tries to cover a wide range of sports classes and to capture the collective activities that are more general in nature. The videos are collected from web resources, indicating that none of them are posed sequences, but rather taken from real-world shootings. In addition, to form an inherently realistic and challenging generalization benchmark, we introduce a novel evaluation protocol called *unseen sports* evaluation, where the training and test splits consist of videos of disjoint sport categories. A robust recognition model that is trained with the collective activities in a certain context, are expected to yield accurate classifications when same collective activities take place in a previously unseen context, *i.e.* in a different sports environment. When the training and test sequences come from different sports, it forms a more realistic and challenging test-bed for evaluating the generalization ability of collective activity recognition.

One of the most interesting aspects of C-Sports dataset is that it is multi-task in nature. One can try to predict the collective activity and at the same time, the sports category label. This dataset provides such a testbed, investigating whether the sports category and collective activity prediction can be carried out simultaneously, and whether the recognition of one task is likely to benefit from the other.

To set the benchmark on C-Sports, we experiment with several state-of-the-art action recognition methods which are representatives of the latest lines of research on this topic. More specifically, we follow three fundamental strategies; i) Two-stream ConvNets [4], where RGB and optical flow representations are explored in conjunction. ii) ConvNet+LSTM-based approach, where spatial information is extracted via ConvNets, and temporal patterns are modeled via LSTMs, and iii) 3D-ConvNets, where spatial and temporal patterns are encoded using 3D convolutions [5]. Each of these models has their own strengths and weaknesses from the recognition point of view. We also introduce the multi-task versions of the 3D-ConvNet [5] and Two-stream ConvNet [4] approaches, which are shown to yield increased performances.

To sum up, the main contributions of this work are as follows:

- We introduce a multi-task dataset called C-Sports, for collective activity and sports category recognition.
- We experiment with several state-of-the-art action recognition methods to set the benchmarks on this dataset.
- We show that the multi-task learning strategy yields significant recognition performance increase; hence suggesting that sport category recognition and collective activity recognition can benefit from each other.
- We provide a new evaluation protocol to assess the generalization ability of collective activity recognition across different sports categories.

Our experimental evaluations demonstrate that using multi-task learning yields promising results, indicating that there is shared knowledge between tasks. The evaluations over unseen sports also indicate that the presence of context influences the recognition performance dramatically; we argue that there is a need for corresponding testbeds to assess whether it is the essence of the activities that is being recognized or other contextual elements. Experiments on the newly introduced evaluation protocol for this purpose, demonstrates that, whilst the standard supervised learning yields high recognition performances, there is still a large room for improvement to recognize collective activities across different sports categories.

2. Related Work

Human action recognition is addressed by hundreds of recent works in computer vision literature (for a detailed survey, please see [6]), yet collective activity recognition problem is relatively a new topic that is less explored. Collective activity recognition is closely related to coherent motion detection in crowd scenes [7], group detection[8] and group activity recognition[9], but these are considered separately because of their difference in problem domains. Collective activity recognition techniques can be grouped into two subcategories: a) shallow approaches and b) deep learning techniques. Below, we give a brief overview of these categories.

2.1. Shallow Approaches

In one of the earliest works, Choi et al. [2] presented a local spatio-temporal descriptor that captures spatial distributions of pedestrians along with their pose over time. With a latent variable framework Lan et al. [10] focus on two new types of interactions, *i.e.* person-person and person-group, and propose adaptive structures for inferring them. Choi and Savarese [3] look at the correlation between motion and activity by means of a hierarchy of activity types that jointly tracks people in a crowd, identifies individual activities together with interactions and resulting the collective activity. Khamis et al. [11] propose combining per-frame and per-track cues. Choi and Savarese [12] create a model by first proposing a descriptor which gathers behaviors of a group of individuals in a spatial-temporal manner to form top-down evidence. Bottom-up information coming from a fragment of tracks and detections are combined with top-down evidence coherently.

Following these studies, recent work shifted focus to capture the relation between a spatio-temporal pattern of each person and their interactions with the crowd. Antic and Ommer [13] use samples of local and group parts gathered by considering their functions and visual similarities, then propose to use max-margin instance learning to train an activity classifier. Tran et al. [14] introduce a graph-based framework for clustering, where edge weights of the graph show how much each person is in interaction with each other.

Amer et al. [15] propose another architecture called Hierarchical Random Field(HiRF), to model higher-order temporal dependencies by only considering dependency hierarchy of model variables. Amer et al. [16] addresses the problem of multi-scale activity recognition, where a three-layered AND-OR graph is proposed to model group activities, actions of individuals, and object participations. To this end, Amer et al. [16] created a new high-resolution video dataset, from UCLA courtyard. Followingly, Amer et al. [17], a Sum-Product Network (SPN) that consists of a mixture distribution of BoWs is used to capture the activity of interest. Although these methods are expensive to implement, they reached state of the art on benchmarks as well as on the Volleyball Dataset[1].

2.2. Deep Approaches

Early works on action recognition and video classification tasks use only CNNs[4, 18], whereas more recent studies Donahue et al. [19] use CNNs with recurrent neural network (RNN) models. In this context, [4, 20, 21] use CNNs in a two-stream manner, where optical flow inputs are used together with RGBs.

Deng et al. [22] proposed a combination of hierarchical graphical models to capture individual actions where a multi-step message passing approach was used between neural network layers. Deng et al. [23] combines graphical models and deep neural networks into a joint framework, where sequential inference is done by a RNN. Hajimirsadeghi et al. [24] presents a Multiple Instance Learning (MIL) framework which uses cardinality relations between latent labels.

In a more recent study, Ibrahim et al. [1] use multiple hierarchical LSTMs, where the first LSTM is used to capture individual actions and the second LSTM is used to evaluate temporal group activity dynamics. This model is further improved in [25] by adding an energy layer instead of a softmax layer. Qi et al. [26] propose an attentive semantic recurrent neural network (RNN), called as stagNet, which uses the spatiotemporal attention and semantic graph to recognize collective activity. A two-level attention based network, person and

Figure 1: Example frames from the eleven sports categories of C-Sports dataset. From top-left to bottom-right, sports categories are *American football, basketball, dodgeball, football, handball, hurling, ice hockey, lacrosse, rugby, volleyball, waterpolo*.

scene levels, is presented by [27] for modeling relationships in group activity recognition. This model modified two-stage Gated Recurrent Units (GRUs) networks to handle temporal variability and consistency. Tang et al. [28] propose a consistency constrained graph model that models the relevant movements of individuals by reducing the importance of irrelevant ones. Tang et al. [29] use the information obtained from the semantic domain for recognizing the collective activities in the training stage of the appearance domain. Zhang et al. [30] present a weakly supervised method which jointly learns an actor detector and collective activity classifier for getting the person-group interaction in scenes. Lu et al. [31] design a graphical convolutional neural network, which investigates interaction relationships in collective activities. Recently, [32] has proposed a multi-stream spatio-temporal convolutional neural network which focuses on person regions in both temporal and spatial channels.

3. Collective Sports (C-Sports) Dataset

3.1. An Overview of C-Sports

C-Sports dataset has been formed out of sports videos, since the sports is one of the most vivid domains that involve collective activities. In addition, sports videos are more easily accessible through the Internet video sharing sites; hence, collecting non-posed video sequences are relatively easier.

During the category selection phase, we have examined many sports categories, and among those, we select the ones that have more tendency to collectivity, where many samples of collective activities are available with relatively higher visual quality. Candidate sports classes are selected from those that have videos comprising least two of the collective activities. For example, football sport class has passing and attacking collective activities. We keep this rule of thumb for collective activity selection as well, i.e., the candidate collective activity must be observable in at least two sports classes.

In C-Sports dataset, there are 11 sports categories and five collective activity categories. Sports categories are *American football, basketball, dodgeball, football, handball, hurling, ice hockey, lacrosse, rugby, volleyball* and *water polo*, whereas five collective activities are *gathering, dismissal, passing, attack* and *wandering*. *Gathering* can be defined as people approaching each other for a specific purpose. *Dismissal* is the separation of people to different directions after gathering. *Pass* is the act of passing items, such as balls, hockey rubbers, etc., between players, whereas *attack* is the movement of the team players towards a specific goal. *Wandering* activity, on the other hand, can be defined as the free movements of team players.

Sample video frames for sports classes are given in Fig. 1, and for collective activity classes in Fig. 2, respectively. Each video in the dataset has two labels, one indicating the sports category and the other indicating the class of the ongoing collective activity.

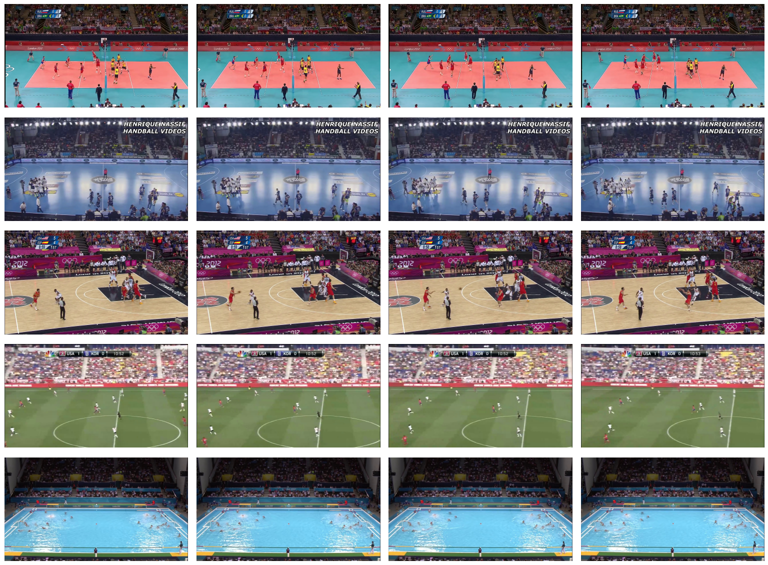

Figure 2: Sample frame sequences of the collective activities of C-Sports dataset. From top to bottom, collective activities are *gathering, dismissal, passing, attack and wandering*.

Table 2: The number of videos and frames in the train/val/test splits of C-Sports in standard supervised evaluation protocol.

	Train	**Validation**	**Test**	**Total**
# videos	1317	435	435	2187
# frames	101300	33228	33407	167935

3.2. Pre-processing

In order to collect videos, several text queries containing the collective activity and the sport class names are formed, and executed on YouTube. Long videos are cropped manually to a range of 5-10 seconds delineating the collective activities. The length is limited to a maximum of 100 frames; ensuring that the start and end of the activities are included within the clip.

After the videos are cleaned and clipped to a certain range, dense optical flows are computed using the method of [33]. The horizontal and vertical components of the displacement vector fields are stored as two optical flow images for a given pair of consecutive frames.

3.3. Evaluation Protocols and Dataset Statistics

Protocol 1 - Standard: First protocol is the standard supervised evaluation protocol; where all the video data is split into disjoint train/validation/test sets and each split includes instances from each sport classes. In this setup, 60% of the video clips are used for training, 20% for validation and 20% for test. The number of videos/frames for this setup are given in Table 2.

Protocol 2 - Unseen Sports: Second evaluation protocol, includes a more interesting and challenging setup, where the training and test are performed on different sport classes. Here, the idea is to assess the generalization ability of models in collective activity recognition task. For this purpose, we divide the dataset based on the sports classes, such that the train and test sets include disjoint sports , *i.e.* the train/test splits do not share any common sports class. Formally, let $\mathcal{Y} = \{1, \ldots, C_a\}$ denote the collective activity classes, and $\mathcal{L} = \{1, \ldots, C_s\}$ denote the set of sports classes. Each training video x_i is annotated with both a collective activity class label y_i and a sports class label l_i. In the unseen sports protocol, in each split, we hold out a subset $\mathcal{L}^u \subset \mathcal{L}$ of sports classes for evaluation purposes as unseen sports classes. Therefore, the training dataset $\mathcal{D}_{train}$ consists of training examples (x_i, y_i, l_i) such that $y_i \in \mathcal{Y}$ and $l_i \in \mathcal{L} \setminus \mathcal{L}^u$, and the test dataset $\mathcal{D}_{test}$ consists of examples (x_j, y_j, l_j) such that $y_j \in \mathcal{Y}$ and $l_j \in \mathcal{L}^u$. In this evaluation protocol, the task is to predict the collective activity class label y_j. To create the splits, we use a cross-validation (CV) approach; there are 11 folds, where each fold corresponds to a particular sports class. In each iteration of CV, 10 folds (*i.e.* all the collective activity videos from 10 sports classes) are used in training and the remaining fold (collective activity videos of the remaining sports class) is used in test. More specifically, all of the collective activities (gathering, dismissal, attack, pass, wandering) are trained on 10 out of 11 sports contexts (*e.g.* American

Football, Dodgeball, Handball, etc.) and tested on one left-out sports context (*e.g.* Water Polo) in each iteration of the cross-validation scheme. The task is to assess whether the recognition models learnt over a set of sports contexts can accurately recognize the collective activities in previously unseen contexts. In this way, it will be possible to evaluate whether the essence of the collective activity is learned independent of the context.

Note that, this is not a zero-shot learning setup; in training, the models have access to all the collective activity classes with a certain set of sports contexts, whereas test is carried out for the same set of collective activity classes on new sports contexts that are not seen during training. When a video from an unseen sports class is encountered, we expect a robust recognition model to recognize the ongoing collective activity, even if it has not seen how that collective activity is carried in the context of the new sport.

In Table 3, the number of videos/frames per each sports classes/collective activity, together with the corresponding totals, are given. According to these statistics, in each split for the unseen sports evaluation, one row of Table 3 is used for test, and the remaining rows are used for training. Note that, some of the cells in this table are zero, indicating that there are no examples found for that particular activity/sports pair. This is mainly due to the unavailability of those actions, for example in water polo, the players do not usually gather or dismiss during match.

4. Methods

In order to provide benchmarks for the newly introduced dataset, we employ three state-of-the-art action recognition models that are built upon recent powerful deep learning strategies: i) ConvNets with LSTM [19], ii) two-stream ConvNets[4], and iii) 3D-ConvNets [5]. In this section, we briefly describe these architectures and also introduce the multi-task versions of the two of the best performing ones. We discuss how these architectures can be utilized for multi-task learning of collective activities and/or sports categories.

Table 3: Dataset statistics based on collective activity and sports classes. Each cell of the table is defined in X/Y format in which X specifies the number of videos and Y denotes the number of frames. Rows correspond to sports classes, whereas the columns correspond to collective activities. In unseen sports evaluation protocol, in each iteration of the cross-validation scheme, the training is carried out over 10 of the rows, where the remaining row is spared for test.

	Gather	Dismissal	Pass	Attack	Wander	**Total**
A.Football	84/5580	11/920	60/3708	71/5146	38/2993	264/18347
Basketball	14/1290	10/960	36/2378	38/2920	44/3670	142/11218
Dodgeball	13/1200	13/1143	57/3625	81/5541	46/3762	210/15271
Football	11/891	13/1036	65/6076	50/4820	13/1300	152/14123
Handball	10/960	15/1470	23/1476	34/2438	29/2815	111/9159
Hurling	10/729	10/810	64/4174	59/4253	50/4134	193/14100
Ice Hockey	13/1106	10/809	48/3169	53/4098	44/3692	168/12874
Lacrosse	109/9222	87/7391	34/2192	45/3303	34/3034	309/25142
Rugby	11/984	11/1017	50/3099	59/4251	47/3786	178/13137
Volleyball	100/7576	99/7872	50/3289	0/0	47/3837	296/22574
Waterpolo	0/0	0/0	72/4812	50/3644	42/3534	164/11990
Total	375/29538	279/23428	559/37998	540/40414	434/36557	2187/167935

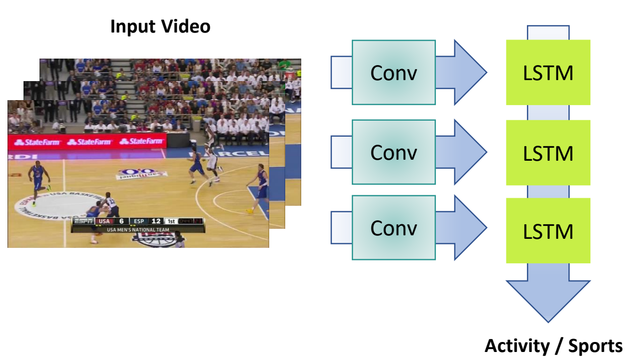

Figure 3: ConvNet + LSTM [19] model architecture.

4.1. ConvNet+LSTM

Thanks to the state-of-the-art results on image classification, it has been appealing to use convolutional neural networks (ConvNets) in video classification with certain adaptations. Karpathy et al. [34] experimented on different techniques to advance connections of ConvNets to make better use of local spatio-temporal information. However, temporal information is likely to get lost if pooling is applied to the extracted features. In order to model the long-range temporal information more adequately, LSTMs Hochreiter and Schmidhuber [35] have been utilized in the literature. Since such a base architecture, *i.e.* using CNNs followed by LSTMs, is frequently used for modeling video content, we choose this model to be the first model to experiment on C-Sports dataset.

Donahue et al. [19] are amongst the first to introduce ConvNets+LSTMs idea and their method is called Long-Term Recurrent Convolutional Networks (LRCN). We adopt their method in our benchmarking. In this architecture, ConvNet layers are used to extract features, and a stack of LSTMs is used to support variable-length sequence prediction. This main idea is illustrated in Fig. 3. We use the model shared by [19], where the LSTM cells are identified with the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{3}$$

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where t denotes the timestep, x_t is input vector at timestep t, h_t is hidden state, W is weight matrix of associated gate, b is bias term. f_t denotes forget gate, i_t is input gate, o_t is output gate, c_t is memory cell state. $\odot$ denotes the element-wise product of vectors at each side. σ is sigmoid function, $tanh$ is hyperbolic tangent function.

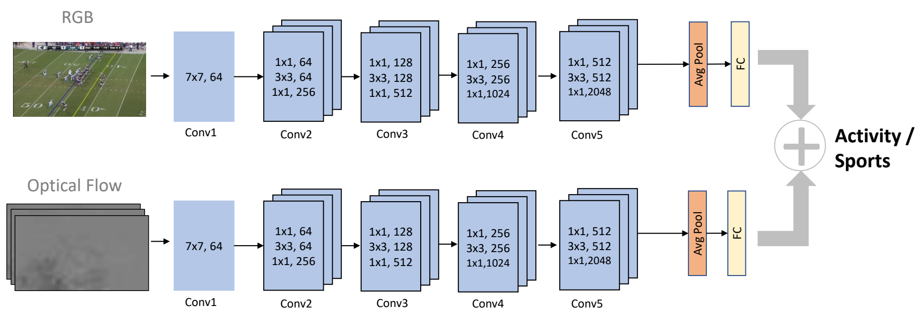

Figure 4: Two-Stream ConvNet architecture.

4.2. Two-Stream Networks

Another dominant idea in modeling video content is the two-stream networks, initially proposed by the seminal work of [4]. Here, the main idea is to use two different CNN streams operating on RGB and optical flow separately, and then to fuse both streams at later stages of the network. The RGB stream is used to capture the spatial information, whereas the optical flow stream is used to capture important movement information across the frames. As stated in [4], static appearances that are associated with particular objects (i.e., basketball, football, hockey rubber) can be used as a clue extracted with the first stream. The second stream, using optical flow as input, brings the crucial information of movement. The two streams have almost identical architecture except for the first convolutional layer. Both of the streams uses the initial weights pre-trained on ImageNet [36]. Figure 4 demonstrates the two-stream model.

4.3. 3D Convolutional Neural Networks

The third model that is frequently used for modeling video sequences is 3D ConvNets. 3D ConvNets are effective for spatio-temporal feature learning via 3-dimensional convolutions and they are commonly used for video modeling and analysis in recent works [37],[38],[39].

The base model of 3D ConvNets is demonstrated in Fig. 5. In 3D ConvNets, instead of the 2D convolutions, 3D kernels are used and convolution is done along 3 dimensions including the temporal dimension. In 2D ConvNets, pooling and

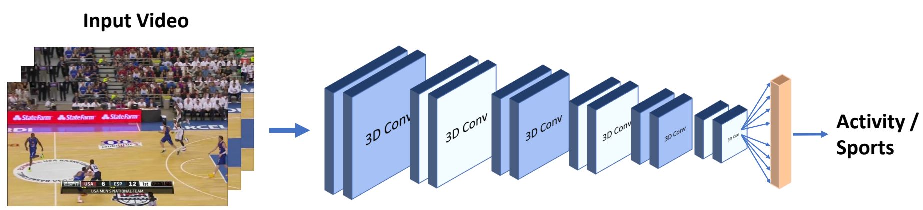

Figure 5: 3D ConvNet architecture.

convolution operations are performed only spatially, whereas in 3D ConvNets, these operations are performed both temporally and spatially. The main idea is to capture temporal information of videos more naturally with 3D convolution and pooling. 3D convolution operation can be formulated as follows [40]

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{7}$$

where $tanh(.)$ is the hyperbolic tangent function, b_{ij} is the bias term of the feature map, m indexes over the set of feature maps in the $(i-1)th$ layer connected to the current feature map, w_{ijm}^{pqr} is the $(p,q,r)th$ value of the kernel connected to the mth feature map in the previous layer, P_i and Q_i are the height and width of the kernel and R_i is the size of the 3D kernel.

Since convolutions are done in 3D, there are more parameters to estimate and therefore, 3D ConvNets are harder to train, requiring a larger volume of data. In our experiments, we have used the base C3D model proposed by [5]. Basically, in this model there are 49 3D convolutional layers, 2 pooling layers and 1 fully connected layer. Batch normalization is used after all convolutional layers and cross-entropy loss is chosen as the loss function.

4.4. Multi-Task Learning (MTL)

Multi-Task Learning (MTL) is a learning method in which multiple learning tasks are solved consequently using a shared model that is learnt jointly. Mathematically, there are k learning tasks T_i for $i = 1, ..., k$ and each task is associated with n_i training samples $\{(x_j, y_j^i)\}_{j=1}^{n_i}$, where x_j is the jth training

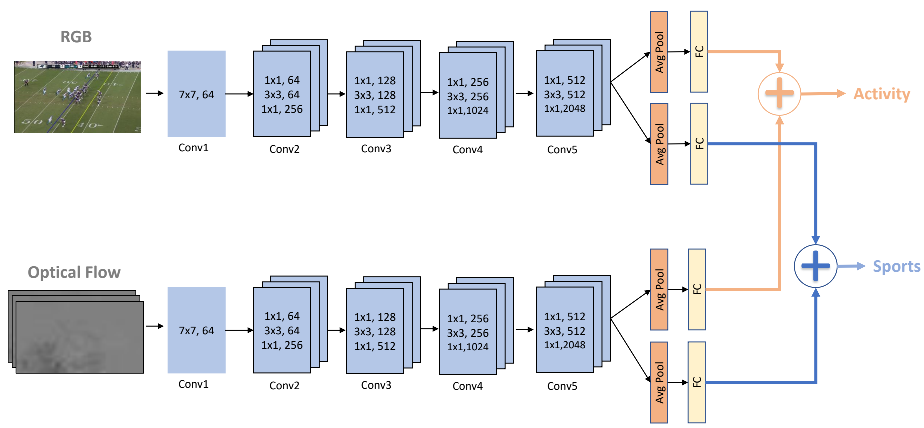

Figure 6: Multi-task version of Two-Stream ConvNet architecture.

instance in T_i and y_j^i is its label. So, there are n_i pairs of data instances and label for each ith task.

In this work, we aim to simultaneously estimate the collective activities and sports categories via a joint estimation model. In order to learn these tasks together, we jointly train our model using multiple loss functions. Specifically, since each problem is a multi-class classification task, we use cross-entropy loss function for each task. In this context, $\mathcal{L}_{act}$ represents the loss function for collective activity recognition as follows

$$\mathcal{L}_{act} = -\sum_{j}^{a} y_j \, log \, (\frac{e^{o_j}}{\sum_i e^{o_i}}) \tag{8}$$

where a represents the number of activity classes, o_j is the output score of the jth collective activity class, and y_j represents the ground truth score of the given class. Similarly, the loss for the sport category recognition, denoted with $\mathcal{L}_{sport}$, is defined as

$$\mathcal{L}_{sport} = -\sum_{j}^{s} y_j \, log \, (\frac{e^{p_j}}{\sum_i e^{p_i}}) \tag{9}$$

where where s represents the number of sports class, p_j is the output score of the jth sports class, and y_j represents the corresponding ground truth score.

The total loss $\mathcal{L}_{total}$ is computed as the equal-weighted sum of these tasks'

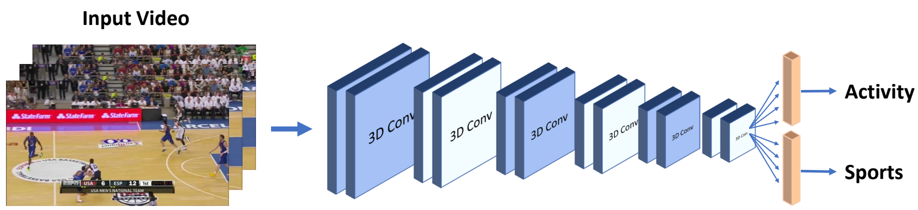

Figure 7: The Multi-task version of 3D ConvNet architecture.

losses as

$$\mathcal{L}_{total} = \mathcal{L}_{act} + \mathcal{L}_{sport}. \quad (10)$$

In order to evaluate the effect of multi-task learning on our tasks, we introduce multi-task versions of both two-stream networks and 3D ConvNets. In both of these methods, we use hard parameter sharing [41] that is applied by sharing layers between all tasks while holding several task-specific output layers. In MTL version of the two-stream network, FC layer in each stream is trained separately. Then, the softmax scores of each stream are added together. This architecture is illustrated in Fig. 6.

The multi-task version of 3D ConvNet is illustrated in Fig. 7. In this method, all layers until the last convolutional layer are trained jointly for both tasks. After the last convolutional layer, we modify the network to have two separate fully connected (FC) layers and two separate softmax classification layers. These FC and classification layers for each stream is trained independently, where a single joint loss is optimized during backpropagation.

5. Experimental Evaluation

5.1. Implementation details

For the ConvNet+LSTM model, we use the same architecture with [19]. The model is based on CaffeNet [42]. Initial ConvNets are trained on UCF101 dataset, then LRCN models are fune-tuned on C-Sports dataset. In the two-stream ConvNet, we use ResNet-50 architecture pre-trained on ImageNet [36]. For the motion stream, we adapt the pre-trained weights on ImageNet [36] as

Table 4: Comparisons of baseline model accuracies using the standard supervised evaluation. The top part shows the performances of the single-task learning methods. The bottom part shows the results of the multi-task versions.

	Architecture	Collective Activities	Sports
single task	ConvNet+LSTM [19]	43.6	70.6
	3D_ConvNet [5]	51.5	85.0
	Spatial ConvNet[4]	30.5	80.4
	Temporal ConvNet[4]	69.2	95.8
	Two_Stream [4]	76.5	98.3
multi-task	3D-ConvNet-MTL	72.6	98.3
	Spatial-MTL	25.7	78.1
	Temporal-MTL	**81.3**	97.4
	Two-Stream-MTL	80.5	**99.0**

well and duplicated these weights to coincide with the 20 channels of optical flow. Then, both streams are fine-tuned using the C-Sports dataset. For 3D-ConvNet, we utilize the ResNet-50 model pre-trained on Kinetics dataset [43], and fine-tuned on the C-Sports dataset.

For Two-Stream and 3D-ConvNet architectures, we train the models for 100 epochs with a learning rate 0.01 and batch size 32. For ConvNet+LSTM model, we train the model for 40K iterations with an initial learning rate 10^{-4} and then decrease the learning rate by a factor of 10 at every 5K iterations. All models are trained on a 12 GB NVIDIA TitanX GPU.

5.2. Results and Discussions

In this section, we present the experimental evaluations of the benchmark methods over C-Sports dataset.

5.2.1. Results with the standard evaluation protocol

As described in Section 3.3, the standard evaluation protocol tests the regular supervised classification case, where examples from each class are available both in training and test. Table 4 shows the results using this standard protocol. Here, the Spatial ConvNet and Temporal ConvNet denote the single ConvNet streams that operate over RGB and optical flow inputs independently, where both ConvNets have ResNet50 architectures. Similarly, Spatial-MTL and Temporal-MTL corresponds to the multi-task versions as described in Section 4.4. Upper part of Table 4 compares the different single-task techniques, whereas the lower part of the table presents the multi-task versions.

When we look at the comparisons in Table 4, for collective activity recognition, we see that the best performance is achieved by the two-stream networks [4], whereas the Temporal ConvNet [4] yields the second best recognition performance. While the Spatial ConvNet produces much lower results compared to the temporal counterpart, we can say that it still includes complementary information, since the fusion of the two streams yields superior performance. The ConvNet+LSTM [19] seems to achieve relatively less successful results amongst the three baseline architectures, whereas the 3D-ConvNet [44] performs comparably better than ConvNet+LSTM approach.

In Table 4, we observe that MTL versions have better recognition performance when compared with the corresponding single task learning (STL) versions, except for Spatial ConvNet. The MTL version of the temporal ConvNet, Temporal-MTL yields the overall best performance for the collective activity recognition, with an accuracy of 81.3%. Another observation is that, 3D-ConvNet method benefits largely from the introduction of the multi-task learning component; where the accuracy has increased more than 20% in collective activity recognition.

Table 4 includes the results for sports category recognition as well. We observe that the recognition rates are higher for sports category recognition. The ConvNet+LSTM[19] method performs considerably lower for this task, whereas,

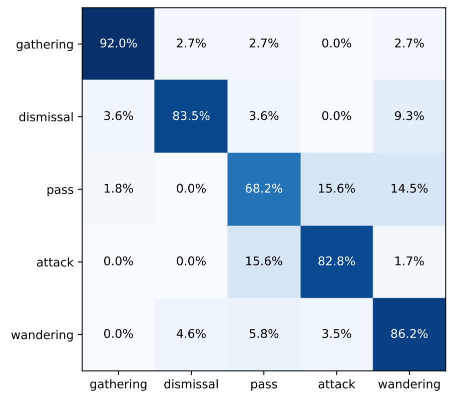

Figure 8: Confusion matrix for the C-Sports dataset using the Temporal-MTL for supervised evaluation protocol. Here, the rows represent the true classes, whereas the columns represents the predicted classes.

similar to collective activity recognition, the introduction of the multi-task learning component raises the recognition ratios for this task as well, yielding a surprisingly high rate of 99% for Two-stream-MTL approach. This suggests that the two tasks, i.e., collective activity recognition and sports category recognition have complementary elements; and joint training of these two tasks is beneficial for the recognition of both.

Figure 8 presents the confusion matrix for the Temporal-MTL method, which achieves the highest accuracy 81.3% for collective activity recognition. According to this matrix, most confusion occurs between the `pass` and `attack` classes. Similarly, a large amount of confusion occurs between the `pass` and `wandering` classes.

In order to investigate further into these confusions, we present the t-SNE visualization of the 3D ConvNet features of the dataset in Fig. 9. In this t-SNE visualization, we observe that the data points are quite scattered; indicating the difficulty of the dataset. Moreover, we observe that the `pass` and `attack` example sequences appear closer in tSNE space. Another observation is that `wandering` class examples are largely scattered, explaining the higher likelihood

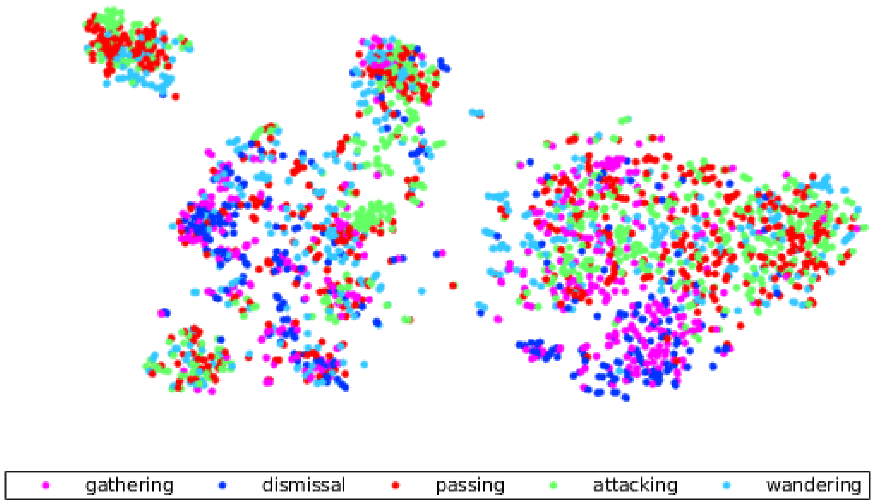

Figure 9: The t-SNE [45] visualization of 3D ConvNet features on the whole C-Sports dataset.

of confusion with other classes.

Fig. 10 shows the class-based accuracies of the MTL based methods in collective activity recognition. Amongst all classes, `wandering` class has the highest recognition ratio, whereas `pass` class has the lowest recognition rates. In four out of five classes, two-stream MTL method performs significantly better, and for `pass` activity, the 3D-ConvNet method yields more successful performance.

5.2.2. Results on Unseen Sports Protocol

While the above supervised evaluation protocol is used as the standard means of evaluation in many collective activity recognition studies, the results may not reflect the generalization ability of the trained classifiers to new types of videos. In order to assess this ability, we introduce another evaluation protocol called *unseen sports* protocol (details given Section 3.3), in which the recognition models trained on a set of sports videos are test on videos of other sport classes, *i.e.* unseen sports classes. The rationale behind this assessment

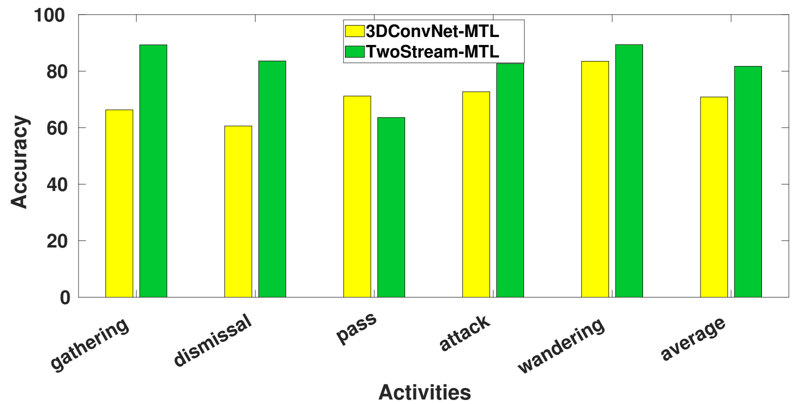

Figure 10: Class-based accuracies of the 3D-ConvNet-MTL and Two-Stream-MTL methods on collective activity recognition using standard supervised evaluation protocol.

is to test whether the classifiers are indeed capturing the essence of collective activities, rather than being largely influenced by the context of the sports.

Since 3D-ConvNet-MTL and Two-Stream-MTL methods produce the most successful results in standard evaluation protocol, we test these two approaches for unseen sports evaluation. Fig. 11 shows the comparative results of 3D-ConvNet-MTL model with the Two-Stream-MTL model. The results are in accordance with the findings in the supervised setting, that the Two-Stream-MTL approach yields significantly superior results compared to 3D-ConvNet-MTL, achieving an accuracy of 59.9% on average.

The individual class accuracies of Two-Stream-MTL method are presented in Table 5. In this table, the columns represent sports classes; rows represent the collective activity classes. For example, the first cell of Table 5 indicates that for recognizing the gathering activity of American Football videos, Two-Stream-MTL model yields 69.0% accuracy when trained on videos of sports classes other than American Football.

According to the results in Table 5 and class-wise average results presented

Table 5: Class-based accuracies of Two-Stream-MTL model using unseen sports evaluation protocol. Each row represents the test results with the corresponding sports videos, when trained on the videos of the rest of the sports classes.

	Gathering	**Dismissal**	**Pass**	**Attack**	**Wandering**	**Avg**
A. Football	69.0	27.2	20.0	91.5	86.8	58.9
Basketball	71.4	70.0	22.2	86.8	90.9	68.2
Dodgeball	15.3	76.9	21.0	0.0	76.0	37.8
Football	54.5	69.2	46.1	94.0	84.6	69.7
Handball	90.0	53.3	52.1	97.0	79.3	74.3
Hurling	40.0	40.0	64.0	89.8	62.0	59.1
Ice Hockey	92.3	60.0	35.4	98.1	59.1	68.9
Lacrosse	97.2	83.9	44.1	66.6	70.5	72.5
Rugby	72.7	54.5	52.0	66.1	95.7	68.2
Volleyball	49.0	54.5	18.0	N/A	87.2	41.7
Waterpolo	N/A	N/A	50.0	60.0	85.7	39.1
Average	59.2	53.6	38.6	68.2	79.8	59.9

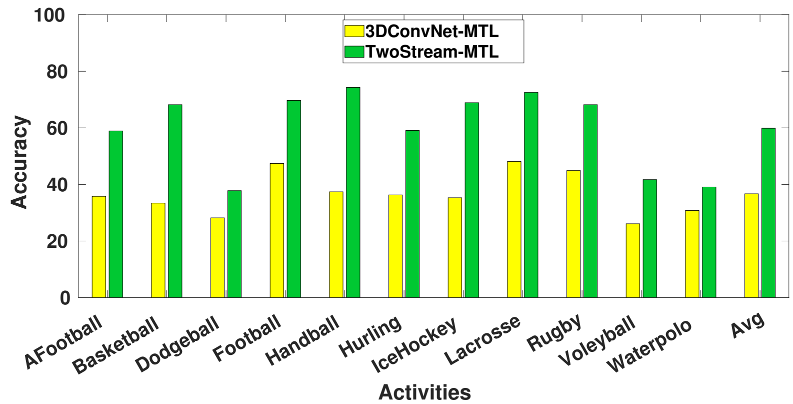

Figure 11: Class-based accuracies of the 3D-ConvNet-MTL and Two-Stream-MTL methods on collective activity recognition using unseen sports evaluation protocol.

in Fig. 11, for individual sports classes, higher accuracies are observed for "Football" and "Lacrosse" classes, which are visually close to each other. On the contrary, lower results are observed for the "WaterPolo" and "Volleyball" classes, which have less visual similarity with the other classes.

When we investigate the results in Table 5 columnwise regarding collective activity classes, we observe that recognition accuracies of `attack` and `wandering` classes are relatively higher, since the visual properties of these classes across different sports do not vary much. We can also say that some of the collective activities, such as `pass`, have a more sport-specific nature, meaning that passing in American Football is quite different than passing in Volleyball; therefore, it is relatively harder to generalize the classifiers for such activities.

Gathering activity is especially recognizable in the Handball, Ice Hockey and Lacrosse sports, yielding impressive recognition results over 90% using Two-Stream-MTL approach. Dismissal activity is mostly recognizable in the Lacrosse videos (83.9% accuracy). The overall accuracy for the `pass` activity is the lowest. One noticeable issue is with the recognition of the `attack` activity in the

Dodgeball videos, where both Two-Stream-MTL and 3D-ConvNet-MTL models completely fail. This may be due to the significant difference in movement direction of the Dodgeball's attack. While the attack activity of the Dodgeball is in the vertical plane, the attack activities in all the other sports classes are carried out in the horizontal plane.

On average, we observe that the recognition results in this "unseen sports" evaluation protocol is significantly lower than the supervised evaluation. In supervised evaluation, the average accuracy of Two-Stream-MTL model is 80.5%, whereas it is 59.9% for unseen sports evaluation. This difference verifies that the recognition models are indeed affected by the surrounding context and they are inclined to fit to context information rather than the essence of collective activities. C-Sports unseen sports evaluation protocol provides a setup for such an evaluation on collective activity recognition domain. We believe that this is an issue that needs further attention from the research community; since collective activity recognition may not be the only domain in which such an phenomenon is likely to happen.

6. Conclusion

In this paper, we present a new benchmark collective activity dataset, called "Collective Sports (C-Sports)", which includes collective activities of sports. The dataset is multi-task in nature; opening up interesting directions to explore. In order to set the benchmarks in this dataset, we experiment with several state-of-the-art sequence recognition approaches in the literature, such as Two-Stream ConvNets and 3D-ConvNets. We also introduce the multi-task versions of the 3D-ConvNet and Two-Stream ConvNet architectures and demonstrate that the multi-task learning improves the recognition accuracies for both collective activity recognition and sports class recognition tasks. To estimate the generalization ability of collective activities more promptly, we introduce a new evaluation protocol that evaluates the recognition models on unseen sports categories. The experimental results on this protocol indicates that the generalization of certain collective activities may be quite limited and this issue remains as an open problem that needs further attention.

To contribute research in this direction, all the data and annotations are available to download[1], together with the extracted optical flow features and trained models.

Acknowledgment

We would like to sincerely thank Emre Boran, Leyla Helin Cetin, Canberk Esmeliler and Hasan Serefal for their help in data collection and annotation. This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) Research Program (1001), Project No 116E102.

References

[1] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, A hierarchical deep temporal model for group activity recognition., in: CVPR, 2016.

[2] W. Choi, K. Shahid, S. Savarese, What are they doing?: Collective activity classification using spatio-temporal relationship among people, in: EEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), IEEE, 2009, pp. 1282–1289.

[3] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in: European Conference on Computer Vision, Springer, 2012, pp. 215–230.

[4] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Adv. Neural. Inf. Process. Syst., 2014, pp. 568–576.

[1] https://cemilzalluhoglu.github.io/csports

[5] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6546–6555.

[6] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: A survey, Image and Vision Computing 60 (2017) 4–21.

[7] W. Lin, Y. Mi, W. Wang, J. Wu, J. Wang, T. Mei, A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes, IEEE Transactions on Image Processing 25 (2016) 1674–1687.

[8] F. Solera, S. Calderara, R. Cucchiara, Structured learning for detection of social groups in crowd, in: Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on, IEEE, 2013, pp. 7–12.

[9] W. Lin, H. Chu, J. Wu, B. Sheng, Z. Chen, A heat-map-based algorithm for recognizing group activities in videos, IEEE Transactions on Circuits and Systems for Video Technology 23 (2013) 1980–1992.

[10] T. Lan, Y. Wang, W. Yang, G. Mori, Beyond actions: Discriminative models for contextual group activities, in: Adv. Neural. Inf. Process. Syst., 2010, pp. 1216–1224.

[11] S. Khamis, V. I. Morariu, L. S. Davis, Combining per-frame and per-track cues for multi-person action recognition, in: European Conference on Computer Vision, Springer, 2012, pp. 116–129.

[12] W. Choi, S. Savarese, Understanding collective activitiesof people from videos, in IEEE Trans. Pattern Anal. Mach. Intell 36 (2014) 1242–1257.

[13] B. Antic, B. Ommer, Learning latent constituents for recognition of group activities in video, in: European Conference on Computer Vision, Springer, 2014, pp. 33–47.

[14] K. Tran, A. Gala, I. Kakadiaris, S. Shah, Activity analysis in crowded environments using social cues for group discovery and human interaction modeling, Pattern Recognition Letters 44 (2014) 49–57.

[15] M. R. Amer, P. Lei, S. Todorovic, Hirf: Hierarchical random field for collective activity recognition in videos, in: European Conference on Computer Vision, Springer, 2014, pp. 572–585.

[16] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, S.-C. Zhu, Cost-sensitive top-down/bottom-up inference for multiscale activity recognition, in: European Conference on Computer Vision, Springer, 2012, pp. 187–200.

[17] M. R. Amer, S. Todorovic, A. Fern, S.-C. Zhu, Monte carlo tree search for scheduling activity recognition, in: Proc. IEEE Int. Conf. Comput. Vis (ICCV), IEEE, 2013, pp. 1353–1360.

[18] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, in IEEE Trans. Pattern Anal. Mach. Intell 35 (2013) 221–231.

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 2625–2634.

[20] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit., 2015, pp. 4305–4314.

[21] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit., 2016, pp. 1933–1941.

[22] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. Roshtkhari, , G. Mori, Deep structured models for group activity recognition, in: British Machine Vis. Conf. (BMVC), 2015.

[23] Z. Deng, A. Vahdat, H. Hu, G. Mori, Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 4772–4781.

[24] H. Hajimirsadeghi, W. Yan, A. Vahdat, G. Mori, Visual recognition by counting instances: A multi-instance cardinality potential kernel, in: Proc. IEEE Computer Vision Pattern Recognition (CVPR), 2015, pp. 2596–2605.

[25] T. Shu, S. Todorovic, S.-C. Zhu, Cern: Confidence-energy recurrent network for group activity recognition, in: Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit. (CVPR), 2017.

[26] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, L. Van Gool, stagnet: An attentive semantic rnn for group activity recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.

[27] L. Lu, H. Di, Y. Lu, L. Zhang, S. Wang, A two-level attention-based interaction model for multi-person activity recognition, Neurocomputing 322 (2018) 195–205.

[28] J. Tang, X. Shu, R. Yan, L. Zhang, Coherence constrained graph lstm for group activity recognition, IEEE transactions on pattern analysis and machine intelligence (2019).

[29] Y. Tang, J. Lu, Z. Wang, M. Yang, J. Zhou, Learning semantics-preserving attention and contextual interaction for group activity recognition, IEEE Transactions on Image Processing (2019).

[30] P. Zhang, Y. Tang, J.-F. Hu, W.-S. Zheng, Fast collective activity recognition under weak supervision, IEEE Transactions on Image Processing (2019).

[31] L. Lu, R. Yu, H. Di, L. Zhang, S. Wang, Y. Lu, Gaim: Graph attention based interaction model for collective activity recognition, IEEE Transactions on Multimedia (2019).

[32] C. Zalluhoglu, N. Ikizler-Cinbis, Region based multi-stream convolutional neural networks for collective activity recognition, Journal of Visual Communication and Image Representation 60 (2019) 170–179.

[33] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: European Conference on Computer Vision, Springer, 2004, pp. 25–36.

[34] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: CVPR, 2014.

[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–80.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009, ieee (2009).

[37] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2010) 221–231.

[38] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: Proceedings of the 11th European Conference on Computer Vision: Part VI, ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 140–153.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[40] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (2012) 221–231.

[41] R. Caruna, Multitask learning: A knowledge-based source of inductive bias, in: Machine Learning: Proceedings of the Tenth International Conference, 1993, pp. 41–48.

[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[43] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, 2015.

[45] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.