Check for updates

Received: 5 June 2022 | Revised: 1 July 2022 | Accepted: 3 July 2022

DOI: 10.1002/cav.2107

SPECIAL ISSUE PAPER

WILEY

Distillation of human–object interaction contexts for action recognition

Muna Almushyti[1,2] | **Frederick W. B. Li**[1]

[1]Department of Computer Science, Durham University, Durham, UK

[2]Deanship of Educational Services, Qassim University, Buraydah, Saudi Arabia

Correspondence
Muna Almushyti, Computer Science Department, Durham University, Upper Mountjoy, South Road, Durham DH1 3LE, UK.
Email: muna.i.almushyti@durham.ac.uk

Funding information
N8 Research Partnership and EPSRC, Grant/Award Number: EP/T022167/1; Unversities of Durham, Manchester and York

Abstract

Modeling spatial-temporal relations is imperative for recognizing human actions, especially when a human is interacting with objects, while multiple objects appear around the human differently over time. Most existing action recognition models focus on learning overall visual cues of a scene but disregard a holistic view of human–object relationships and interactions, that is, how a human interacts with respect to short-term task for completion and long-term goal. We therefore argue to improve human action recognition by exploiting both the local and global contexts of human–object interactions (HOIs). In this paper, we propose the Global-Local Interaction Distillation Network (GLIDN), learning human and object interactions through space and time via knowledge distillation for holistic HOI understanding. GLIDN encodes humans and objects into graph nodes and learns local and global relations via graph attention network. The local context graphs learn the relation between humans and objects at a frame level by capturing their co-occurrence at a specific time step. The global relation graph is constructed based on the video-level of human and object interactions, identifying their long-term relations throughout a video sequence. We also investigate how knowledge from these graphs can be distilled to their counterparts for improving HOI recognition. Finally, we evaluate our model by conducting comprehensive experiments on two datasets including Charades and CAD-120. Our method outperforms the baselines and counterpart approaches.

KEYWORDS

global context, graph attention network local context, human–object interaction

1 | INTRODUCTION

Human action recognition tasks typically involve interaction with objects. Such tasks are challenging even for deep learning methods especially under complex scenarios. A human can interact with the same object but performing different actions. For example, a human can hold a laptop and can put it somewhere. These two actions, "hold" and "put," are different but they involve the same object. In addition, a variety types of objects afforded to same action

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Computer Animation and Virtual Worlds* published by John Wiley & Sons Ltd.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(e.g., refrigerators and doors can be involved in the same interactions including open and close) needs to be considered.[1] Moreover, the existence of different objects around a human could confuse model predictions. For example, if a human is drinking a coffee and there is a book nearby, a model may inaccurately predict that the human is both reading and drinking. Furthermore, during a video sequence, the states of humans and objects change over time, such as a human can hold an object and release it at any time step, followed by interacting with another object which makes identifying correct interactions very challenging. Hence, identifying humans and objects at each time step and learning their relations can help understand a scene. This implies learning objects that are closely located for identifying interactions. The transition of human and object states over time also offers crucial cues for understanding what a human is performing. Consequently, it is important to capture contextual information about interactions at a specific time and throughout a video, making action recognition success. Although modeling human–object interactions (HOIs) has been broadly studied in images,[2-5] it has received less consideration in videos. Even deep learning methods have been developed for recognizing human actions in videos, most of them, including Covnet,[6] recurrent neural networks (RNNs)[7,8] and 3D convolution models,[9,10] only take individual frame-wise information as inputs without explicitly modeling human–object relations across a video sequence. Hence, such methods failed to capture useful global context cues, that is, long-term human–object dependencies, for assisting action recognition.

Recent works[11-15] have proposed to model human–object relations by performing spatiotemporal reasoning through multihead attention mechanism for recognizing actions in videos. As they capture more context cues to reason HOIs, they have achieved promising results over baselines that do not consider human–object relations.

In this work, we propose to capture human–object relations from their local and global views as well as transferring knowledge between these views. The local view captures human–object relations at a specific time, for example, spatial relation. The global view encodes human–object relations over time, for example, temporal relation, to capture long-term human–object relations. The design of the network for global and local views is flexible. Motivated by the success of graph attention networks (GATs)[16] in different tasks including person re-identification,[17] action recognition,[11,18,19] and video question answering,[20] our method exploits GAT to construct our two contextual views modules. Since the global context of an interaction offers complementary information to the local contexts of such interaction and vice versa, previous works combined different types of context features via concatenation[14] or summation,[11] or even considered the global features as an extra node in the graph.[21] Inspired by Reference 22 and instead of learning these contexts via features level which are prone to noise, we propose to apply knowledge distillation, transferring knowledge about interactions from global to local views, and vice versa. We therefore exploit teacher-student network design, investigating which of the proposed contextual views can form a better teacher, offering richer HOI information to guide the student network for improving action recognition performance.

To the best of our knowledge, we are the first to investigate knowledge distillations between two HOI views for action recognition in videos. Our main contributions are:

- Proposing a novel teacher–student network based on graphs neural networks to learn spatial and temporal interrelations between humans and objects in a video from two different contextual views. Hence, long-term and nonlocal dependency between human and objects across video frames can be captured.
- Investigating how knowledge from the teacher contextual view of interactions can be obtained, and distilling it to the student view of interactions to improve action recognition performance.
- Evaluating our model on Charades and CAD-120[23] datasets[24] and conducting comprehensive experiments in transferring knowledge between local (e.g., Spatial) and global (e.g., Temporal) context views of HOIs. Our teacher–student design is effective to distill knowledge between global context and local context graphs. We also observe that the student network outperforms its teacher by exploiting both global and local contexts of an interaction.

2 | RELATED WORK

2.1 | Action recognition models in videos

The simple models for action recognition can be done by extracting frame features through CNNs followed by pooling via averaging, or followed by RNNs to model the sequence of frames for predicting actions in videos.[7,25] Recently, space-time

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

models are proposed, such as 3D convolutions. They add an extra time dimension to kernels in order to extract spatiotemporal features from videos.[26-29] Likewise, I3D model[9] has been introduced by inflating pretrained 2D convolution kernels to 3D for extracting space-time features from video clips. In addition, X3D network[30] expands 2D architecture across other axes including depth, spatial, width, and frame rate which enable training the network with fewer parameters than other 3D networks such as Slowfast, yielding comparable results. There are related methods focusing on long-term dependency as in Reference 31 where the temporal relations between frames at different time scales are modeled.[31,32] Moreover, structural temporal modeling has been proposed, which uses a two-level modeling approach to capture both short- and long-term temporal information.[33,34] Nonlocal relations between pixels in space and time are also studied for recognizing actions in videos.

More recently, transformer-based frameworks such as Reference 35 are proposed for recognizing actions in videos where the transformer is used to get discriminative features from each frame and then being aggregated via attention. Transformer is also used for action recognition networks purely without utilizing convolutions.[36] In addition, beside the appearance features that can be extracted from RGB images, optical flow and depth data are used to enhance human action recognition in videos.[6,37-40]

The above works focus on whole video features rather than on important cues of an action such as interobjects or interhuman relations that our method considers. Also, our method only focuses on visual information from videos to model HOIs for action recognition.

2.2 | Spatiotemporal reasoning for action recognition

Spatiotemporal reasoning involves detecting humans and objects and modeling their relations to capture contextual information for classifying an action. In Reference 41, the relation between objects at specific time and the objects from adjacent frames at specific window is learned via Feature Bank Operator, such as nonlocal, to capture long-term context in videos. Inspired by the success of RNNs in modeling sequence data, such as Long Short-Term Memory (LSTM), they have also been used for spatiotemporal reasoning over objects in videos.[15] Space and time graphs have been proposed in Reference 11, where object context relations during time is captured and objects in adjacent frames are connected based on their intersection over unions (IOU). A relation network is proposed to focus on the relation between actors and video-level features for identifying actions.[42] To capture high order object interactions, attention mechanism is applied over objects at each frame followed by a LSTM process.

Furthermore, in Reference 43, graph attention is used to model the relations between human and objects, considering their spatial distance in each clip. Transformers are also used in learning visual relations between the features of humans located in the centre clip, which is considered as a query, and the features from the whole clip in order to learn the context of the action by using the properties of self-attention in the transformer.[44] Our work proposes to use two different contexts of human and object relations, capturing different cues of an interaction that helps recognize actions. Inspired by References 16,17, we choose GAT as a base network for learning such interactions.

2.3 | Knowledge distillation

Distilling knowledge has been proposed to transfer knowledge learned from ensemble of classifiers or large network into a small network.[45] This implies compressing complex networks without losing their performance.[46] It can be done by minimizing the loss between small network (student) predictions and the large network's soften labels (teacher). Recently, Knowledge distillation (KD) is extended and combined with privileged information,[47] where additional information is available only during training time to form a generalized distillation.[48] For action recognition task, the knowledge is distilled between multiple modalities (e.g., skeleton, RGB, optical flow), which can be considered as privilege information and not all of them are available during inference.[49-52] Also, KD is employed in different directions, such as defencing against adversarial attacks,[53] classifying unlabeled data via unifying diverse classifiers.[54] To increase segmentation accuracy, KD has also been applied to semantic segmentation, for example, by distilling intraclass feature variation or interclass distance from teacher network to student.[55,56] Furthermore, KD is used for improving object detectors by selecting different valuable areas (e.g., foreground) to distill.[57-59] Inspired by these directions, we extend it to HOI recognition in videos, allowing knowledge transfer between global and local contextual views of interactions via KD.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3 | GLOBAL–LOCAL INTERACTION DISTILLATION NETWORK

3.1 | Network overview

Figure 1 shows the architecture of our GLIDN. It takes video frames and the bounding boxes of human and objects at each frame as inputs. Frame features (e.g., appearance features) are then extracted by a convolutional neural network, such as ResNet.[60] RoIAlign[61] is then applied to extract features of each human and object boxes from the backbone feature map. The bounding boxes are generated via RPN[62] if they are not available in the dataset. These extracted region features are used as the initial features of graph nodes in both the global and local contextual views. The human–objects relations from the teacher view are distilled into the student context representation by aligning logits from the two contextual views.

3.2 | Global and local context graphs

As mentioned earlier, we utilize GAT[16] as our graph networks to learn the relations between human and objects from different contextual views.

The global context graph is constructed to learn the relation between each entity (e.g., human or object) and all other entities in a video. The graph is constructed based on the learned adjacency matrix between humans and objects over time in a video as in Reference 11. Hence, the interaction score between two nodes in GAT is:

$$\alpha_{i,j} = \sigma(a[W_o(x_i)|W_o(x_j)]), \tag{1}$$

where W_o is a learnable transformation which is shared between object nodes in a video. a is a weight matrix projecting the concatenated features to a scalar that reflects attention coefficient between two nodes (e.g., humans or objects). "|" indicates concatenation. In this global context graph, coefficients represent the learned interaction scores between humans and objects. In other words, $\alpha_{i,j}$ is a scalar that represents the relation between two nodes i and j (e.g., edge) in the adjacency matrix A, which is of the size $N \times N$ where N is the number of humans and objects that appeared in the video. σ is a nonlinearity function such as LeakyReLU. Later, $\alpha_{i,j}$ is normalized across all other nodes within the video with respect to node i via softmax. Thus, the updated node features via GAT can be formulated as:

$$x_i = \sum_{j \in N} \alpha_{i,j} W_o x_j. \tag{2}$$

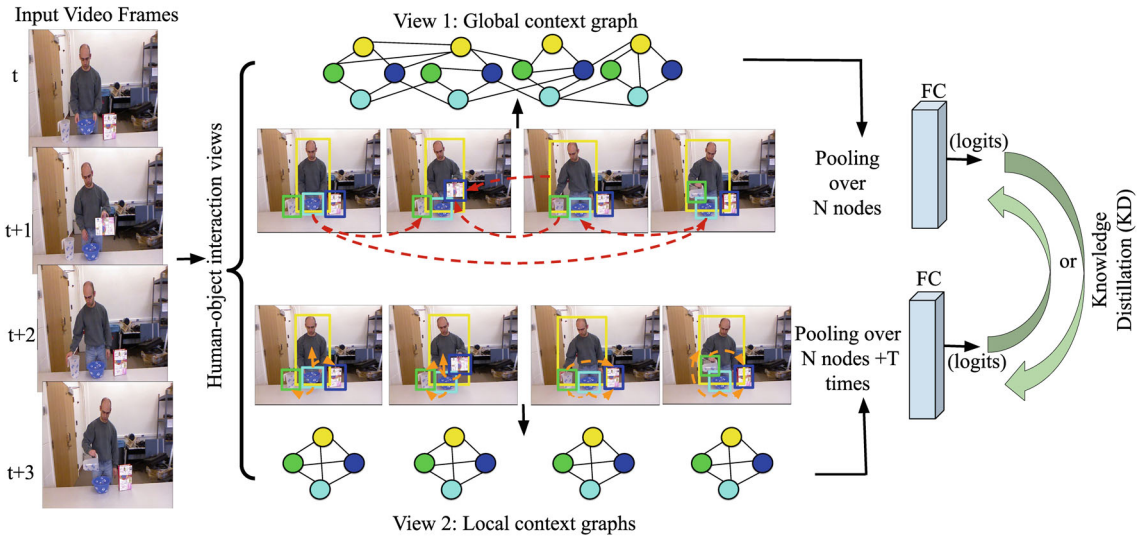

FIGURE 1 Overview of our proposed global–local interaction distillation network.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Through this graph, long-term dependency of HOIs in a video can be captured since each object is attended to all other objects over the video at different time frames.

On the other hand, in the local context, there are T number of graphs, where T indicates the number of frames in the video. Through these local graphs, besides relations induced by closely located humans and objects, nonlocal dependency relations between human and objects in a video frame can also be captured. Nonlocal means when objects and humans are distant from each other within a frame. Hence, each node captures local contextual information via learning relation with other nodes (e.g., human or objects) within the same frame regardless they are spatially close to or distant from each other. Local context is therefore learned from various interactions in which humans / objects attend to others in the same frame.

In short, the way of updating graph nodes is the same in both global and local graphs using Equation (2), yet the nodes relation scope is different. In global graph, each graph node attends (learns relation) to all other nodes in the video. In contrast, in local graph, only relations between nodes at the same frame is learned. Hence, the local and global contexts use the same operation (e.g., GAT) but consider different structures. Through these graphs, the relations between humans and objects can be learned even though they are not nearby in space and time. Hence, various human–object, object–object and human–human relations within individual frames and throughout a video can be extensively learned.

3.3 | Global and local context distillation

In order to have an informative representation of HOIs, features from both global and local contextual views should be fully utilized. This may not be simply done by combining features from the two contexts, despite it is a standard way for gathering information from different sources or views. In contrast, we adapt a teacher–student framework to utilize global and local context of HOIs through knowledge distillation. To implement such a knowledge transfer, we incorporate soft labels from the teacher context graph network to guide the student context graph network during training. These soft targets are probability distributions from the logits in the teacher network.

In our experiments, different distillation losses are utilized, depending on the nature of a dataset. For CAD-120 dataset, we minimize the KL divergence between soften labels of teacher and student as in References 22,63. For Charades, we use l_2 loss as distillation loss to meet the property of training multilabel videos. Hence, the l_2 distillation loss can be formulated as:[64]

$$L_{\text{Distill}} = \frac{1}{n}\sum_{i=1}^{n}(P(t)_i - P(s)_i)$$
$$P(s)_i = \frac{1}{1+e^{\frac{l_c}{T}}}, \tag{3}$$

where $P(t)_i$ and $P(s)_i$ are softened sigmoid predictions from teacher and student networks, respectively. l_c is the logit from the last fully connected layer in the network, and T is a hyperparameter that represents the temperature for class c.[64]

3.4 | Training

We first train teacher network, which captures one context view (e.g., global context) of HOIs along with hard labels, using cross-entropy loss. We then fix the teacher network and train the student network which is another view of HOIs (e.g., local context). Hence, the objective function for training the student network can be:

$$L_{\text{student}} = \lambda_1 L_{\text{CE}} + \lambda_2 L_{\text{Distill}}, \tag{4}$$

where L_{CE} is cross-entropy loss between student predictions and hard labels (e.g., ground truth). λ_1 and λ_2 are hyperparameters for balancing the two losses and are set empirically (see Section 4.4). For testing, the results is reported using only the student network.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 | EXPERIMENTS

4.1 | Datasets and settings

4.1.1 | Datasets

We conduct extensive experiments on two public datasets, including Charades[24] and CAD-120.[23] We particularly choose these datasets not only because they are used for evaluating action recognition models but also because they have a variety of human object interactions where our paper focuses on. We demonstrate the flexibility and capability of modeling human interactions via our proposed model by considering large-scale and small datasets as well as diverse 2D and 3D backbones.

Charades dataset[24] consists of 9848 multilabel videos with indoor daily activities that involve humans interacting with various types of objects. The number of videos in training phase is about 8K videos and 1.8K for validation. There are 157 action classes in total. Figure 2 shows some HOI examples in Charades dataset.

Moreover, CAD-120[23] contains 120 videos where 10 different daily life interactions are performed by four different subjects. Depth images and skeleton information are available besides RGB frames but we use only the RGB images.

4.1.2 | Evaluation metric

Since Charades dataset is a multilabel video dataset, we use mean average precision to report the final results. In contrast, each video in CAD-120[23] has only one activity label. Thus, accuracy is adopted as the evaluation metric as in Reference 65.

4.2 | Implementation details

4.2.1 | Charades dataset

For training our GLIDN, we follow training procedure in Reference 11 and we use Inflated 3D ConvNet (I3D) model[9] with Resnet-50 and Slowfast-R50[29] as our backbone networks. In I3D backbone, we initialize it with pretrained parameters on

FIGURE 2 Examples of human–object interactions from Charades dataset.[24]

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 A summary of training settings in our experiments on CAD-120[23] and Charades.[24]

Dataset	Optimizer	LR	Epochs	Decay	Number of GAT Layers	Training procedure
CAD-120[23]	Adam	2.e-5	100	Each 50 steps	3	Leave-One-Out Cross-Validation
Charades[24]	SGD	0.018	60,30	Each 40 steps	1	Stage-Wise Training (two stages)

Kinetics-400 dataset[66] from Reference 67. For Slowfast-R50 backbone, we adopt it from Reference 67 where it is already trained on Charades dataset. We sample 32 and 64 frames as in References 29 and 11) from each video as input with 224 × 224 pixels for I3D and Slowfast-R50, respectively. The inputs are randomly cropped such that the shorter side is sampled in [256, 320] pixels. We train I3D backbone for 60 epochs with a batch size of eight videos, where the learning rate is set to 0.018 for the first 40 epochs and is reduced by a factor of 10 for the last 20 epochs. Following the previous works including,[11-13] we use stage-wise training strategy where the model is trained end-to-end in the second stage for 30 epochs.

As in Reference 11, we apply RoIAlign on the output feature maps of the backbones (before the FC) and each node in the graph is with a fixed dimension of 7 × 7× 512 (1 ×1× 512 via max pooling).

Since Charades dataset does not provide human and object bounding boxes, we use Region Proposal Network (RPN) in Faster R-CNN[62] to produce object proposals. We use the top 15 proposals at each frame. These proposal features (bounding boxes) represent human and object nodes in the graphs.

We adapt binary cross-entropy with sigmoid activation as a loss function for multilabel video classification in addition to the distillation loss.

For inference, we perform multicrop-view inference on each video. In other word, we sample 10 clips from each videos and perform multi-crop testing as in Reference 12. Later, the result is reported based on fusing scores from 30 views via max pooling.

4.2.2 | CAD-120 dataset

We sample 30 frames uniformally from each video and we used the bounding box annotations that are provided within the dataset. We follow Reference 68 for extracting features for human and objects nodes. For each bounding box in a frame, we apply RoI cropping and then reshape it to meet the input size of 224 ×224× 3 for 2D ResNet backbone. Therefore, human and object node features are with the size of 2048 dimension that are produced by ResNet-50.

Besides distillation loss, we train our model with cross-entropy loss with an initial learning rate of 2.e-5. We train our model for 100 epochs in total using Adam optimizer.[69] Our network is trained on a single Nvidia TITAN RTX 24GB GPU. Hyper-parameters for our training are summarized in Table 1. Appendix S1 contains further details.

4.3 | Comparison with state-of-the-arts

As shown in Tables 2 and 3, we compare our GLIDN with all prior methods that applied on CAD-120 and Charades datasets, respectively. Our approach achieves the best performance. It is noted that on Charades, our network outperforms the baselines including I3D and Slowfast, which do not consider spatiotemporal contextual views of objects.

TABLE 2 Accuracy (%) results on the CAD-120 dataset[23]

Model	Accuracy%
Wang et al.[70]	81.2
Liu et al.[71a]	93.3
koppula et al.[23a]	80.6
Tayyub et al.[72a]	95.2
Sanou et al.[65]	86.4
GLIDN (ours)	**92.85**

[a] Prior works make use of additional skeleton or depth information and thus are not directly comparable to our approach.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Classification mAP (%) results on the Charades dataset[24]

Model	Backbone	mAP%
2-Stream[73]	VGG-16	18.6
2-Stream +LSTM[73]	VGG-16	17.8
Async-TF[73]	VGG-16	22.4
a Multiscale TRN[31]	Inception	25.2
I3D[9]	Inception	32.9
I3D[11]	R50-I3D	31.8
STRG[11]	R50-I3D	36.2
STAG[12]	R50-I3D	37.2
Pose and Joint-Aware[74]	R50-I3D	32.81
GLIDN (ours)	R50-I3D	**37.51**
LFB Max[41]	R50-I3D-NL	38.6
Slowfast 16×8[29]	R50-3D	38.9
Slowfast 16×8 + GLIDN **(ours)**	R50-3D	**41.00**

TABLE 4 Comparison of graph node settings with prior works on Charades[24]

Model	Number of nodes	Nodes information	mAP%
STRG[11]	50	Objects	36.20
STRG[11]	25	Objects	35.9
STAG[12]	15	Objects and edges*	37.20
GLIDN **(ours)**	15	Objects	**37.51**

Note: *indicates edges which represent the union box of two object nodes.

Our network also performs better than STRG,[11] which has used spatiotemporal object relations. Although our global context graph is the same as in STRG[11] in term of the temporal range of objects and human, there are three main differences. First, we use graph attention instead of graph convolution network that used in STRG. Second, in our model, we consider this graph as a teacher or a student network whereas in STRG it is just a graph that is combined with another nonlearnable "spatio-tempral graph." Third, we explore knowledge distillation for capturing more HOI contextual cues, while the work in STRG follows the common method for training their model (e.g., binary cross-entropy loss only). This implies that our approach of using different views of object relations via distillation can help the model generalize better in identifying different types of interactions. Thus, our method has achieved better results even with much fewer number of proposals, as shown in Table 4.

Notably, our approach of utilizing two different views of HOIs and their knowledge transfer can offer more informative cues about interaction even without any human–object abstract information (e.g., the union of both objects) as in Reference 12. This indicates the importance of context modeling of humans and objects without the need of additional information (e.g., visual phrases).

Moreover, our choice of GAT for learning human–object relations in both global and local views is important since we have achieved 35.35 comparing to 34.2 in Reference 11 for the global context with fewer number of nodes. Consequently, we have achieved the best results on Charades comparing to prior works that use the same backbone networks.

We have also achieved better results on the CAD-120[23] than other works that use temporal sampling and 3D CNN[65,70] without fine tuning and with the use of object features extracted from 2D backbone. This implies our KD from different views can remarkably contribute to HOIs reasoning, as it can better capture long-term temporal structure of interactions.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

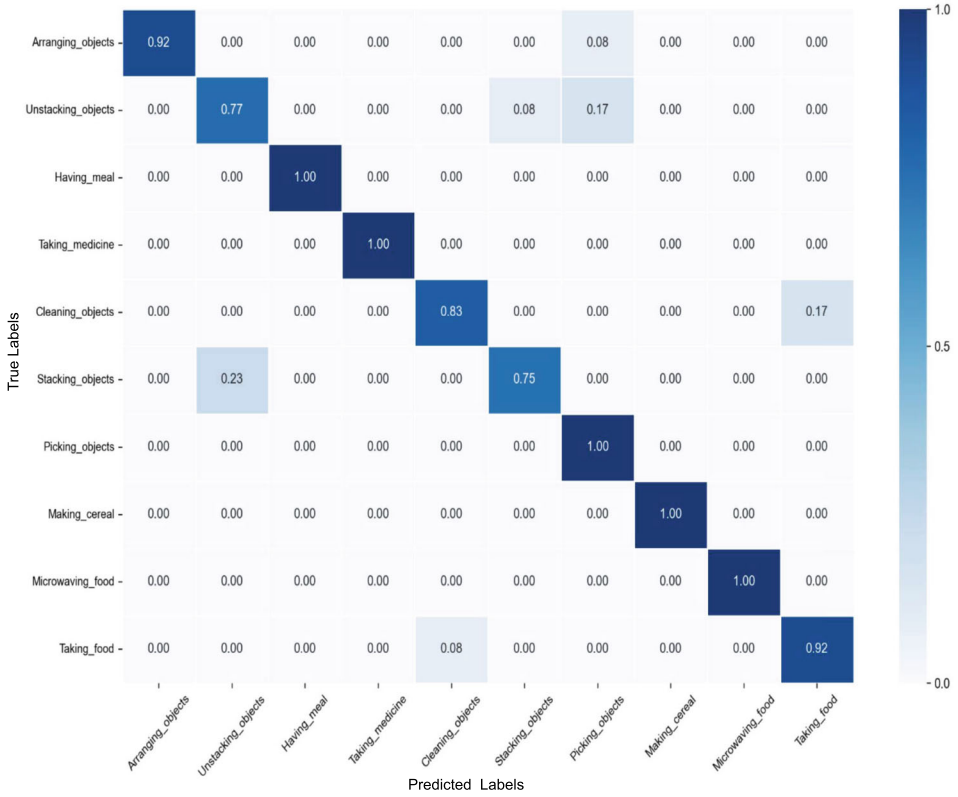

FIGURE 3 Confusion matrix for the CAD-120 dataset[23] when using our proposed global–local interaction distillation network

The confusion matrix in Figure 3 studies how well our method can predict actions correctly based on CAD-120. It can be observed that most false predicted actions relate to stacking and unstacking objects or some actions alike. Such actions usually involve the same object but being different in human movement directions. This may be resolved by capturing more temporal information, such as increasing the number of sampled frames.

4.4 | Ablation studies

To evaluate our proposed GLIDN, we conduct ablation studies to demonstrate the impact of each part of our GLIDN on learning HOIs. We first evaluate the baseline without any of interaction contextual views. We then evaluate our network by using each of the contextual views independently. Finally, we report the performance of our complete network. The ablation study results are shown in Table 5 for Charades[24] and CAD-120 datasets.[23]

TABLE 5 Ablation results the CAD-120[23] and Charades[24] datasets

Model	Charades[24](Slowfast)	Charades[24](I3D)	CAD-120[23](2D R-50)
Baseline	38.9	34.23	74.17
Local-context (spatial)	40.73	36.45	84.97
Global-context (temporal)	39.95	35.39	84.75
Context views fusion (e.g. Concat)	40.43	36.81	85.22
Late Fusion	40.95	37.23	85.97
Local-teacher	39.89	**37.51**	87.76
Global-teacher	**41.00**	36.99	**92.85**

Note: Results from two different backbones are reported on Charades.[24]

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.4.1 | Are contextual views of humans and objects important?

As shown in Table 5, running our network without any human–object relations or with only a single view (either local or global view) degrades the network performance. Clearly, when we consider only human and object information (e.g., via concatenation) without learning their relation, the performance of the network decreases significantly by 14% in CAD-120.[23]

Also, when considering only human-object temporal relations on Charades,[24] the performance drops more than 1% mAP, which reflects the importance of local relations between human and objects at a specific time as they can provide useful context information. This indicates that some of the interactions can be recognized by focusing on the spatial relation, especially with the existence of multiple objects around a human. Finally, capturing both the global and local human-object relations via distillation can help transfer the complementary information from the teacher view to the student contextual view. Hence, the ablation experiments illustrate that each component of the proposed GLIDN plays toward improving the model performance, where 41.00% mAP is achieved on Charades.

4.4.2 | Which of the contextual views play the roles of the teacher network?

In the original KD, the teacher network is larger than the student network. In contrast, in our work, both student and teacher networks give informative cues about interactions from different contextual views. We hence conduct comprehensive experiments to decide which of the contextual view can better serve the teacher role. Logically, when we take into account the wide range of information provided by the global context, we can consider it as a larger view for HOIs since each human/object learns a relation with all other humans/objects throughout all video frames, while the local context only provides information about how humans/objects attend the others within each individual frame. This idea is evaluated on Charades[24] and CAD-120 datasets.[23] As shown in Table 5, best results are usually achieved when we consider the global contextual view as the teacher. Hence, we can conclude that the temporal View (e.g., global contextual view of HOIs) is mostly a viable candidate for the teacher.

However, we notice that utilizing different backbones on the same dataset as in Charades,[24] leading to different selection of teacher network. This suggests that the features retrieved from different backbones have an impact on determining which of the contextual views play a better role as the teacher. For instance, when training our method with I3D backbone on the Charades dataset,[24] we find that using the local contextual view as a teacher achieves better performance. The reason behind this is that the final representation in Slowfast experiments involves concatenating objects relations with fast path features, which are from 64 frames. This means that Slowfast backbone is richer in temporal information than the I3D backbone, which only uses 32-frame features. Hence, when the temporal range is not large enough to capture better contextual information, especially in clutter background videos as in Charades,[24] the spatial local context teacher may outperform the temporal global one. Our findings indicate that distilling the knowledge of interactions between the global and local views outperforms other counterpart approaches in both scenarios, whether a teacher is taking a local or global contextual view.

There are other factors controlling the distillation process, namely the hyper-parameters of T (temperature), λ_1 and λ_2 (weights for balancing the losses in Equation 5). We conduct comprehensive experiments in both CAD-120[23] and Charades[24] using different values of these hyper-parameters. Two forms of λ settings are used for balancing the weight between the two terms of the objective function as in Equation (5). In the first form of setting, we used the generalized distillation form as in Reference 48 where λ_1 is equal to (1- λ_2). The second form is by setting λ_1 to 1 and λ_2 to 4 or 0.7 as shown at the first two rows in Table 6 which shows the results of applying different hyper-parameters on CAD-120 dataset[23] with different settings for teacher and student.

We observe that the best values of T are different for both global contextual view and local view since each network view produces different probability distribution for the logits. We also find in the global teacher, the temperature of 1 achieves the best accuracy as in Reference 75 when the weight λ_2 is equal to 0.7. Moreover, when we consider local contextual view as the teacher network, we observe that a large value of T (e.g., 5) with a distillation weight of 0.3 produces the best result of 87.76%. Hence, the optimal values of T and λ can be set empirically based on the predictions of the teacher network. Appendix S1 provides further hyper-parameters details for Charades Dataset.[24]

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 Accuracy results on CAD-120 dataset[23] after applying different values of T (temperature) and λ_2 (weight of the distillation loss)

T	λ_2	Global–teacher%	Local–teacher%
2	4	87.56	84.36
1	0.7	**92.85**	86.00
5	0.3	88.36	**87.76**
10	0.3	88.45	83.53
20	0.3	87.62	83.50
5	0.5	84.33	86.84
10	0.5	85.69	84.25
20	0.5	87.47	83.59
5	0.7	86.84	81.89
10	0.7	88.54	82.61
20	0.7	85.27	86.00

TABLE 7 Comparison between deep mutual learning (DML) and teacher–student networks for distilling knowledge between object contexts on CAD-120 Dataset.[23]

Model	Accuracy%
DML (local)	87.73
DML (global)	86.64
Our GLIDN (Global–teacher)	**92.85**

4.4.3 | Is teacher–student network design a good choice for distilling object contexts?

In order to evaluate our teacher–student network design, we compare it with other collaborative learning approaches, such as Deep Mutual Learning (DML),[76] where the two contexts views are jointly trained. As presented in Table 7, we can observed that our teacher–student network achieves a better result of 92.85% with an large increase of 6.21% when we consider the teacher network as the global context of HOIs, while 86.64% is achieved via DML. This is because the teacher–student network approach allows the use of contextual information from the teacher network guiding the student network to capture much structural knowledge about HOIs.

4.4.4 | Is context distillation better than conventional fusion?

In order to compare our proposed context distillation for recognizing HOIs with standard methods for combining the features and capturing complementary cues from the two views, we conduct two experiments including early fusion and late fusing methods. In early fusion, we concatenate the features from the two views, then fed them to a classifier. In contrast, for late fusion, we average the predictions of views. As in Table 5, we can observe that our model captures better cues of interactions, whereas in the early fusion, some noise in features may affect the network performance. Moreover, as stated in Reference 22 that knowledge distillation can be considered as a late fusion method, we may confirm this statement in Charades dataset[24] where the model performance via a late fusion is similar to knowledge distillation with only the 0.05% and 0.3% improvement. Although, the late fusion and knowledge distillation results in Charades[24] are close, the results of both approaches outperform the baseline and single view context, proving our claim of exploiting the context of human object interactions from two different views. On the other hand, we find that distilling knowledge between HOI contexts outperforms the late fusion on CAD-120.[23] The late fusion model achieved an accuracy of 85.97%, but when knowledge distillation is applied with the same training setting, the performance is improved by 6.88%. This supports our claim that knowledge distillation can be used to capture the context of HOIs from many views.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 4 Example frames from video ID:0510180218.[23] Bounding boxes are not displayed for clarity.

FIGURE 5 Example frames from video ID:1204144736.[23] Bounding boxes are not displayed for clarity.

4.5 | Evaluation examples

Figures 4 and 5 show two video examples from CAD-120 Dataset.[23] We found from the examples that inconsistent recognition results may be come up if only one contextual view is applied. Since the context of HOIs varies, it is difficult to determine which contextual view is more effective.

For example, in Figure 4, the video is with the correct label "taking food" and it is misclassified as "arranging objects" using only the local contextual view. However, the video is recognized correctly by using the global view where temporal interactions between human and objects are learned at the video level. This indicates the importance of observing the change of object and human status over time, which is captured via the global context.

Another example is shown in Figure 5, where the correct label of the video is "stacking objects" but it is misclassified as "unstacking objects" when using the global contextual view only. However, the video can be correctly classified using the spatial contextual view. This illustrates the importance of having specific time human-object relations, which provides some structure information about an interaction, via its local view.

Notably, our GLIDN model classifies both videos correctly when we consider the global context view (e.g., temporal) as a teacher. This implies that distilling local and global contextual information increases the generalizability of the model. We have achieved the best results of 92.85% on CAD-120 dataset.[23]

5 | EXPLORING THE DESIGN OF THE TEACHER NETWORK

We now investigate alternative designs for distilling human and object (H-O) contexts between different views.

We explore the way of extracting H-O contexts from multiteacher settings. In this design, the spatial graph at each frame acts as a teacher. These frame-based teachers are trained with shared parameters. Teachers in this situation learn human and object spatial relationships in a frame and generate knowledge (e.g., logits) at the frame-level. In contrast,

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 8 Accuracy results on CAD-120 dataset[23] after applying different designs of teachers

Model	Accuracy%
Spatial-Multi-teacher (30 frames)	86.30
Spatial-Multi-teacher (15 frames)	83.49
Spatial-single teacher as in our GLIDN	87.76
Temporal-teacher our GLIDN (S 15 frames)	87.56
Temporal-teacher our GLIDN (S 30 frames)	**92.57**

Notes: S indicates student network. In the last two rows, student network is trained with 15 and 30 frames, respectively.

our GLIDN only has one teacher, which generates predictions based on a video's frame relations. Hence, GLIN teacher performs a video-level prediction. We train the student network, which is the global graph with many teachers from various frames that consider the local relations between human and objects. Hence, the knowledge is distilled from multiple spatial views teachers. The corresponding loss used in training the student can be written as:

$$L_{\text{student}} = \lambda_1 L_{\text{CE}} + \lambda_2 \left(\frac{1}{N} \sum_{n=1}^{N} L_{\text{Distill}_{(S,T^n)}} \right), \tag{5}$$

where N is the number of teachers that participate in the student network's training. Table 8 shows the outcomes of utilizing several instructors by using different samples of frames as teachers (e.g., 30 or 15 frames), while the student network remains the same in both situations (e.g., 30 frames).

As can be observed from the results that considering the spatial relations based on single teacher (e.g., video-level) produces better knowledge, which can be distilled to the student. Also, the temporal teacher outperforms the spatial teachers in both single-teacher and multiple-teacher settings. Furthermore, even with fewer frames, the temporal teacher can still lead the spatial student.

6 | CONCLUSION

The context of HOIs gives crucial cues about how human interacts with different objects. Our GLIDN, a novel human objects interaction distillation network, explicitly uses two different views of humans and objects context to capture their interactions at specific time and throughout a video. We also propose context knowledge distillation to transfer knowledge from the teacher contextual view of HOIs to the student network that has information from different context of such interactions. Extensive experiments demonstrate that we outperforms prior works on two datasets including Charades[24] and CAD-120.[23] Our future work will explore self-supervised approaches for identifying human and objects and their interactions in videos to overcome the need for human and object bounding boxes information, which are not available in most video datasets, while RPN may not accurately detect some objects.

FUNDING INFORMATION

This work made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester and York.

ORCID

Muna Almushyti https://orcid.org/0000-0002-7828-7553
Frederick W. B. Li https://orcid.org/0000-0002-4283-4228

REFERENCES

1. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.
2. Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: crowdsourcing data collection for activity understanding. Proceedings of the European Conference on Computer Vision. Springer; 2016, p. 510–26.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3. Koppula HS, Gupta R, Saxena A. Learning human activities and object affordances from rgb-d videos. Int J Robot Res. 2013;32(8):951–70.
4. Xu B, Wong Y, Li J, Zhao Q, Kankanhalli MS. Learning to detect human-object interactions with knowledge. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019.
5. Chao YW, Liu Y, Liu X, Zeng H, Deng J. Learning to detect human-object interactions. Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2018, p. 381–9.
6. Gkioxari G, Girshick R, Dollár P, He K. Detecting and recognizing human-object interactions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018, p. 8359–67
7. Bansal A, Rambhatla SS, Shrivastava A, Chellappa R. Detecting human-object interactions via functional generalization. arXiv preprint arXiv:1904.03181, 2019.
8. Xu B, Li J, Wong Y, Zhao Q, Kankanhalli MS. Interact as you intend: Intention-driven human-object interaction detection. IEEE Trans Multimed. 2019;22(6):1423-1432.
9. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Adv Neural Inf Process Syst. 2014;27:568–76.
10. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015, p. 2625–34.
11. Li, F., Gan, C., Liu, X., Bian, Y., Long, X., Li, Y., et al. Temporal modeling approaches for large-scale youtube-8m video understanding. arXiv preprint arXiv:1707.04555, 2017.
12. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017, p. 6299–308.
13. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018, p. 6450–9.
14. Wang X, Gupta A. Videos as space-time region graphs. Proceedings of the European Conference on Computer Vision (ECCV); 2018, p. 399–417
15. Herzig R, Levi E, Xu H, Gao H, Brosh E, Wang X, et al. Spatio-temporal action graph networks. Proceedings of the IEEE International Conference on Computer Vision Workshops; 2019.
16. Tan H, Wang L, Zhang Q, Gao Z, Zheng N, Hua G. Object affordances graph network for action recognition. BMVC; 2019.
17. Materzynska J, Xiao T, Herzig R, Xu H, Wang X, Darrell, T.: Something-else: compositional action recognition with spatial-temporal interaction networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020, p. 1049–59.
18. Baradel F, Neverova N, Wolf C, Mille J, Mori G. Object level visual reasoning in videos. Proceedings of the European Conference on Computer Vision (ECCV); 2018, p. 105–21.
19. Yang J, Zheng WS, Yang Q, Chen YC, Tian Q. Spatial-temporal graph convolutional network for video-based person re-identification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020, p. 3289–99.
20. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence; vol. 32, 2018.
21. Lu L, Lu Y, Yu R, Di H, Zhang L, Wang S. Gaim: graph attention interaction model for collective activity recognition. IEEE Trans Multimed. 2019;22(2):524–39.
22. Huang D, Chen P, Zeng R, Du Q, Tan M, Gan C. Location-aware graph convolutional networks for video question answering. Proceedings of the AAAI Conference on Artificial Intelligence; vol. 34, 2020, p. 11,021–8.
23. Ghosh P, Yao Y, Davis L, Divakaran A. Stacked spatio-temporal graph convolutional networks for action segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2020, p. 576–85.
24. Pan B, Cai H, Huang DA, Lee KH, Gaidon A, Adeli E, et al. Spatio-temporal graph for video captioning with knowledge distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020, p. 10,870–9.
25. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: deep networks for video classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015, p. 4694–702
26. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2012;35(1):221–31.
27. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision; 2015, p. 4489–97.
28. Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell. 2017;40(6):1510–7.
29. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019, p. 6202–11.
30. Feichtenhofer C. X3d: expanding architectures for efficient video recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020, p. 203–13.
31. Zhou B, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. Proceedings of the European Conference on Computer Vision (ECCV); 2018, p. 803–18.
32. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, et al. Temporal segment networks for action recognition in videos. IEEE Trans Pattern Anal Mach Intell. 2018;41(11):2740–55.
33. Wang L, Tong Z, Ji B, Wu G. Tdn: temporal difference networks for efficient action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021, p. 1895–904.

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

34. Liu Z, Wang L, Wu W, Qian, C., Lu, T. Tam: temporal adaptive module for video recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021, p. 13,708–18.
35. Neimark D, Bar O, Zohar M, Asselmann D. Video transformer network. arXiv preprint arXiv:2102.00719; 2021
36. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: a video vision transformer. arXiv preprint arXiv:2103.15691, 2021.
37. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H. Skeleton-based action recognition with shift graph convolutional network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
38. Cheng K, Zhang Y, Cao C, Shi L, Cheng J, Lu H. Decoupling gcn with dropgraph module for skeleton-based action recognition.
39. Si C, Jing Y, Wang W, Wang L, Tan T. Skeleton-based action recognition with spatial reasoning and temporal stack learning. Proceedings of the European Conference on Computer Vision (ECCV); 2018, pp. 103–18
40. Tu Z, Li H, Zhang D, Dauwels J, Li B, Yuan J. Action-stage emphasized spatiotemporal vlad for video action recognition. IEEE Trans Image Process. 2019;28(6):2799–812.
41. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R. Long-term feature banks for detailed video understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019, pp. 284–93.
42. Sun C, Shrivastava A, Vondrick C, Murphy K, Sukthankar R, Schmid, C. Actor-centric relation network. Proceedings of the European Conference on Computer Vision (ECCV); 2018, p. 318–34.
43. Tomei M, Baraldi L, Calderara S, Bronzin S, Cucchiara R. Video action detection by learning graph-based spatio-temporal interactions. Comput Vis Image Understand. 2021;206:103187.
44. Girdhar R, Carreira J, Doersch C, Zisserman A. Video action transformer network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019, p. 244–53.
45. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015
46. Thoker FM, Gall J. Cross-modal knowledge distillation for action recognition. Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). IEEE; 2019, p. 6–10.
47. Vapnik V, Vashist A. A new learning paradigm: learning using privileged information. Neural Netw. 2009;22(5-6):544–57.
48. Lopez-Paz D, Bottou L, Schölkopf B, Vapnik V. Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643, 2015.
49. Luo Z, Hsieh JT, Jiang L, Niebles JC, Fei-Fei L. Graph distillation for action detection with privileged modalities. Proceedings of the European Conference on Computer Vision (ECCV); 2018, p. 166–83.
50. Garcia NC, Bargal SA, Ablavsky V, Morerio P, Murino V, Sclaroff S. Distillation multiple choice learning for multimodal action recognition. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021, p. 2755–64.
51. Crasto N, Weinzaepfel P, Alahari K, Schmid C. Mars: motion-augmented rgb stream for action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019, p. 7882–91
52. Dai R, Das S, Bremond F. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2021, p. 13,053–64.
53. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. Proceedings of the 2016 IEEE symposium on security and privacy (SP). IEEE; 2016, p. 582–97.
54. Vongkulbhisal J, Vinayavekhin P, Visentini-Scarzanella M. Unifying heterogeneous classifiers with distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019, p. 3175–84.
55. Zhang Z, Zhou C, Tu Z. Distilling inter-class distance for semantic segmentation. arXiv preprint arXiv:2205.03650, 2022.
56. Wang Y, Zhou W, Jiang T, Bai X, Xu Y. Intra-class feature variation distillation for semantic segmentation. Proceedings of the European Conference on Computer Vision. Springer; 2020, p. 346–62.
57. Zhang L, Ma K. Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors. Proceedings of the International Conference on Learning Representations; 2020.
58. Guo J, Han K, Wang Y, Wu H, Chen X, Xu C, et al. Distilling object detectors via decoupled features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021, p. 2154–64.
59. Yang Z, Li Z, Jiang X, Gong Y, Yuan Z, Zhao D, et al. Focal and global knowledge distillation for detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020, p. 4643–52.
60. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016, p. 770–8.
61. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision; 2017, p. 2961–9.
62. Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst. 2015;28:91–9.
63. Bian C, Feng W, Wan L, Wang S. Structural knowledge distillation for efficient skeleton-based action recognition. IEEE Trans Image Process. 2021;30:2963–76.
64. Liu Y, Sheng L, Shao J, Yan J, Xiang S, Pan C. Multi-label image classification via knowledge distillation from weakly-supervised detection. Proceedings of the 26th ACM International Conference on Multimedia; 2018, p. 700–8.
65. Sanou I, Conte D, Cardot H. An extensible deep architecture for action recognition problem. Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2019); 2019.
66. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.
67. Fan H, Li Y, Xiong B, Lo WY, Feichtenhofer C. Pyslowfast; 2020. https://github.com/facebookresearch/slowfast

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

68. Sunkesula SPR, Dabral R, Ramakrishnan G. Lighten: learning interactions with graph and hierarchical temporal networks for hoi in videos. Proceedings of the 28th ACM International Conference on Multimedia; 2020, p. 691–9.
69. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
70. Wang K, Wang X, Lin L, Wang M, Zuo W. 3D human activity recognition with reconfigurable convolutional neural networks. Proceedings of the 22nd ACM International Conference on Multimedia; 2014, p. 97–106.
71. Liu Z, Yao Y, Liu Y, Zhu Y, Tao Z, Wang L, et al. Learning dynamic spatio-temporal relations for human activity recognition. IEEE Access. 2020;8:130,340–52.
72. Tayyub J, Tavanai A, Gatsoulis Y, Cohn AG, Hogg DC. Qualitative and quantitative spatio-temporal relations in daily living activity recognition. Proceedings of the Asian Conference on Computer Vision. Springer 2014, p. 115–30.
73. Sigurdsson GA, Divvala S, Farhadi A, Gupta A. Asynchronous temporal fields for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017, p. 585–94
74. Shah A, Mishra S, Bansal A, Chen JC, Chellappa R, Shrivastava A. Pose and joint-aware action recognition. arXiv preprint arXiv:2010.08164, 2020.
75. Girdhar R, Tran D, Torresani L, Ramanan D. Distinit: learning video representations without a single labeled video. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019, p. 852–61.
76. Zhang Y, Xiang T, Hospedales TM, Lu H. Deep mutual learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018, p. 4320–8.

AUTHOR BIOGRAPHIES

Muna Almushyti is pursuing a PhD degree in Computer Science at Durham University, Durham, UK. She completed her Master's degree in Computer Science from the School of Electrical Engineering and Computer Science at Ottawa University, Ottawa, Canada in 2014. She works as a lecturer at Computer Department in the Deanship of Educational Services at the Qassim University, Qassim, SA. Her research interests include keystroke dynamics, authentication, Human-object interaction recognition and Generative Adversarial Networks (GANs) in videos.

Frederick W. B. Li is an Associate Professor at Durham University. He received his PhD degree from City University of Hong Kong. He has served as a guest Editor for several special issues of World Wide Web Journal, Journal of Multimedia and JDET. He has also served as a Conference Co-Chair / Program Co-Chair for ICWL conferences and a Program Co-Chair of ISVC. His research interests include computer graphics, collaborative virtual environments and educational technology.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Almushyti M, Li FWB. Distillation of human–object interaction contexts for action recognition. Comput Anim Virtual Worlds. 2022;33(5):e2107. https://doi.org/10.1002/cav.2107

1546427x, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cav.2107 by Durham University - University Library and Collections, Wiley Online Library on [12/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License