Towards Explainable Action Recognition by Salient Qualitative Spatial Object Relation Chains

Hua Hua,[1] **Dongxu Li,** [1] **Ruiqi Li,** [1] **Peng Zhang,**[1] **Jochen Renz,** [1] **Anthony Cohn** [2]

[1] Research School of Computer Science, Australian National University, Canberra, Australia
[2] School of Computing, University of Leeds, Leeds, UK
{hua.hua, dongxu.li, ruiqi.li, p.zhang, jochen.renz}@anu.edu.au, a.g.cohn@leeds.ac.uk

Abstract

In order to be trusted by humans, Artificial Intelligence agents should be able to describe rationales behind their decisions. One such application is human action recognition in critical or sensitive scenarios, where trustworthy and explainable action recognizers are expected. For example, reliable pedestrian action recognition is essential for self-driving cars and explanations for real-time decision making are critical for investigations if an accident happens. In this regard, learning-based approaches, despite their popularity and accuracy, are disadvantageous due to their limited interpretability.

This paper presents a novel neuro-symbolic approach that recognizes actions from videos with human-understandable explanations. Specifically, we first propose to represent videos symbolically by qualitative spatial relations between objects called *qualitative spatial object relation chains*. We further develop a *neural saliency estimator* to capture the correlation between such object relation chains and the occurrence of actions. Given an unseen video, this neural saliency estimator is able to tell which object relation chains are more important for the action recognized. We evaluate our approach on two real-life video datasets, with respect to recognition accuracy and the quality of generated action explanations. Experiments show that our approach achieves superior performance on both aspects to previous symbolic approaches, thus facilitating trustworthy intelligent decision making. Our approach can be used to augment state of the art learning approaches with explainabilities.

Introduction

While learning-based approaches are very popular (LeCun, Bengio, and Hinton 2015), criticisms remain over their reliability and explainability in making high-stake decisions (Rudin 2019). Video action recognition (i.e. recognize which action occurs in a video) is one such typical task. For example, failures in recognizing pedestrian actions by self-driving cars may lead to critical safety issues and those failures could be caused by incorrect reasoning. Most state-of-the-art action recognition approaches (Lin, Gan, and Han 2019; Tran et al. 2019; Carreira and Zisserman 2017; Feichtenhofer et al. 2019; Li et al. 2020a,c,b; Qian et al. 2021) cannot be easily understood or justified by humans as they use neural networks in a black-box fashion (Rudin 2019).

We argue that a video action recognizer is trustworthy or at least its trustworthiness is measurable if it is able to justify its decisions in a human-understandable way. Such justifications can be the existence of certain objects, object relations, or changes in object relations (Zhuo et al. 2019; Li et al. 2021). For example, in soccer matches, *offside* occurs if there is a particular change in positional relations between ball and players.

We will investigate action recognition and explanation by qualitative spatial relations; it is not difficult to extend our approach and relevant discussions to non-qualitative or non-spatial relations. A spatial relation is a relation that is relevant to the spatial properties of objects. For example, connectivity, direction, size, distance, moving direction or moving speed. Spatial relations between objects are essential in many real-world actions and one typical example is shown in Figure 1a. A relation is qualitative if it is a symbol with specific meaning (e.g. "close") rather than a numeric value (e.g. "0.1 meter away from"). We choose qualitative spatial relations for two reasons: first, qualitative spatial relations are intuitive and close to human cognition and thus being human-understandable (Chen et al. 2015); second, it is possible to detect such relations without involving much human labor (Gatsoulis et al. 2016).

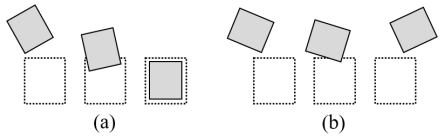

Figure 1: (a) Essential spatial relations when performing the action of *car parking*: at first, the car (grey rectangle) is disconnected from the parking space (dotted rectangle) and then they overlap; in the end, the car is included in the parking space. (b) An accidental interaction between the car and the parking space.

Qualitative spatial relations are formally defined in qualitative spatial calculi (Renz and Nebel 2007; Cohn and Renz 2008). One of the most widely applied qualitative spatial calculi is Region Connection Calculus (RCC) (Randell, Cui, and Cohn 1992), which consists of topological relations (or connectivity relations) between regions (e.g. disjoint or overlapping). The *car parking* example in Figure 1a can be de-

scribed as a chain of objects and their relations: o_1-o_2-dc-po-pp, where o_1 stands for the car (grey rectangle), o_2 stands for the parking space (dotted rectangle), and dc, po and pp are RCC relations that indicate two regions are disconnected, partially overlapping, or included respectively.

In this paper, we aim to achieve *explainable action recognition* by exploring critical qualitative spatial relations or relation changes between objects. The motivation of our approach is to provide explanations for some of the results from "black-box" approaches. Our main contributions are threefold: (1) we introduce a novel symbolic action representation based on qualitative spatial object relation chains (or object relation chains in short); based on this representation, (2) we develop a *neural saliency estimator* that measures the correlation between object relation chains and actions in different contexts; and (3) we generate *action explanations* by identifying salient object relation chains. Based on these innovations, our action recognizer achieves better performance than previous symbolic approaches in both recognition accuracy and the quality of generated action explanations.

Related Work

In this section, we give a brief introduction to previous research on qualitative spatial calculi and attempts attempts to apply them to recognize or explain actions.

Qualitative Spatial Calculi

As mentioned in the introduction, qualitative spatial relations are defined in qualitative spatial calculi. In this subsection, we will first introduce several representative qualitative spatial calculi and then overview qualitative spatial calculi to show that qualitative spatial relations are able to describe spatial relations in different aspects; involving multiple forms of objects; and in various granularities.

Representative qualitative spatial calculi The Region Connection Calculus (RCC) (Randell, Cui, and Cohn 1992) describes the topological relations between regions. Its very original form is RCC-8, which consists of eight relations as shown in Figure 2a. Other RCC variations are with different granularities (e.g. RCC-2, RCC-3, RCC-4, or RCC-5) and their mapping relations are shown in Figure 2b (Gatsoulis et al. 2016). From RCC-8 to RCC-2, topological relations are defined with gradually coarser granularities: RCC-2 only discriminates whether two regions are connected. The $\mathcal{STAR}$ (Renz and Mitra 2004) calculus originates from the Cardinal Direction Calculus (CDC) (Frank 1991; Ligozat 1998) and it consists of qualitative direction relations between points with arbitrary granularity (Figure 3a and 3b). In contrast, Cardinal Direction Relations (CDR) (Skiadopoulos and Koubarakis 2004) focuses on binary direction relations between regions in a specific granularity: the space is divided into 9 tiles with the first region in the middle tile; the relation between the second region and the referencing region is decided by which tile(s) the second region is located in. As shown in Figure 3c and 3d, the Qualitative Distance Calculus (QDC) (Clementini, Di Felice, and Hernández 1997) is also designed with arbitrary granularity and is able to describe qualitative distance relations and the span of each distance relation is flexible and can be either predefined by users or learnt by clustering metric distances, which is similar to the case in (Tayyub et al. 2014) where time durations are clustered into short, equal and long.

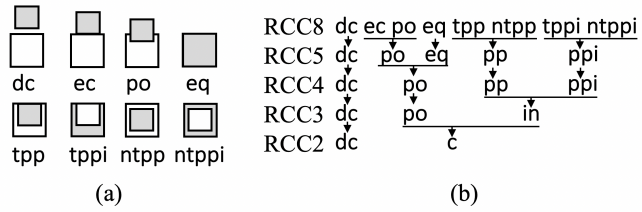

Figure 2: (a) RCC-8 is the initial form of RCC and it consists of eight relations: dc (disconnected), ec (externally connected), po (partially overlapping), eq (equal), tpp (tangential proper part), $tppi$ (tangential proper part inverse), $ntpp$ (non-tangential proper part) and $ntppi$ (non-tangential proper part inverse). (b) The mappings between RCC relations with different granularities. c, pp, ppi are three new relations defined from those mappings: pp is the union of tpp and $ntpp$, ppi is the union of $tppi$ and $ntppi$; and c is the union of po and pp.

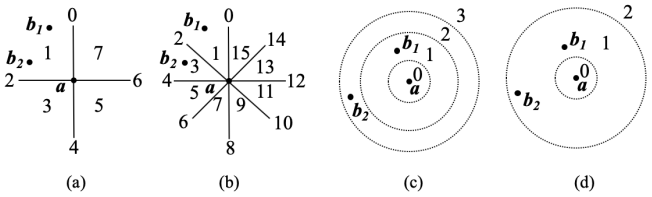

Figure 3: (a) In $\mathcal{STAR}_2$ the 2D space is divided into 8 sectors by 2 lines. The $\mathcal{STAR}_2$ relations between ab_1 and ab_2 are both 1. (b) In $\mathcal{STAR}_4$ the 2D space is divided into 16 sectors by 4 lines. The $\mathcal{STAR}_4$ relation between a and b_1 is 1 while that between a and b_2 is 3. (c) QDC_4: the 2D space is divided into 4 sectors by 3 circles. The QDC_4 relation between a and b_1 is 1 while that between a and b_2 is 2. (d) QDC_3: the 2D space is divided into 3 sectors by 2 circles. The QDC_2 relation between ab_1 and ab_2 are both 1.

Spatio-temporal aspects There are different taxonomies of spatio-temporal aspects in (Renz and Nebel 2007; Dylla et al. 2017). Based on both of them, there are at least five essential spatio-temporal aspects: direction, distance, motion, size, and topology. Time is regarded as a special and implicit aspect because it has to be considered in almost all spatio-temporal actions (as long as there are changes in spatial relations). Most qualitative spatial calculi only cover one aspect: the $\mathcal{STAR}$ Calculus (Renz and Mitra 2004) focuses on qualitative direction relations between points with arbitrary granularity (Figure 3a and 3b); the Qualitative Distance Calculus (QDC) (Clementini, Di Felice, and Hernández 1997) is also designed with arbitrary granularity and is used to describe qualitative distance relations as in Figure 3c and 3d; the Qualitative Trajectory Calculus (QTC) (Van de Weghe et al. 2006) models the moving direction and speed relations between objects; Point Algebra (PA) (Van Beek 1992) (which

consists of three relations $>$, $<$ and $=$) can be used to describe the relative size of objects; RCC-m consists of only topological relations as in Figure 2. There are also q-models that cover multiple aspects. For example, the Ternary Point Configuration Calculus (TPCC) (Moratz and Ragni 2008) is a combination of direction and distance relations and CORE-9 (Cohn, Renz, and Sridhar 2012) is relevant to multiple spatio-temporal aspects: direction, topology, etc.

Object forms and relation arities Qualitative spatial calculi that are relevant to the same aspect can vary because of different abstract forms of objects and relation arities. For example, the $\mathcal{STAR}$ Calculus consists of binary direction relations between points; Cardinal Direction Relations (CDR) (Skiadopoulos and Koubarakis 2004) focuses on binary direction relations between regions; while the Double Cross Calculus (DCC) (Freksa 1992) deals with ternary direction relations between points.

Granularities Some qualitative spatial calculi have different variations to be flexible with different granularities. For example, the granularities of $\mathcal{STAR}_m$ and QDC$_m$ can be different positive integers (as shown in Figure 2). A special case is the RCC family: there are at least five variations of RCC-m with different granularities. Namely, m can be $2, 3, 4, 5,$ or 8. As shown in Figure 4b, coarser variations are generated from finer variations by mapping multiple finer relations into a single coarser relation or dropping some of the finer relations. For example, as shown in Figure 2, two RCC-3 relations po and pp are mapped into a RCC-2 relation c (which indicates two regions are connected) (Gatsoulis et al. 2016). The granularity of a qualitative spatial calculus can be decided by users or learnt by clustering metric relations, which is similar to the case in (Tayyub et al. 2014) where numeric time durations are clustered into short, equal and long. For example, in (Galata et al. 2002), numeric direction, velocity, and distance relations between cars are clustered into qualitative spatio-temporal relations, which is the basis of describing high-level interactive behaviors between cars. Similar technique is also utilized in (Bleser et al. 2015).

Qualitative Spatial Action Description

In this subsection, we will introduce previous work that utilizes qualitative spatial relations to describe actions. In (Dubba et al. 2015), actions like airplane arrival or departure are represented as changes in the RCC relations between airplanes, vehicles, and ground zones. Such representations are then used to recognize actions in unseen videos (Dubba et al. 2015). In (Tayyub et al. 2014), the occurrence of various object relation chains like o_1-o_2-dc, o_2-o_3-dc-po, o_1-o_3-dc-po-dc has been counted and used as features of machine learning approachs to recognize daily activities. A similar but more sophisticated representation strategy is applied in (Sridhar, Cohn, and Hogg 2010; Duckworth, Hogg, and Cohn 2019) to achieve unsupervised action clustering. In (Young and Hawes 2015), four qualitative spatial calculi RCC, $\mathcal{STAR}$, QDC, and QTC have been applied to describe the qualitative movement states of agents in a RoboCup soccer simulator. Their qualitative spatial relation based system models soccer player actions like kick or dash.

In most previous work, the set of candidate qualitative spatial relations is decided by human knowledge. Namely, most previous researchers manually determine which set of qualitative spatial calculi is chosen to describe actions. An experimental study has been conducted in [Sridhar et al., 2011] where the performance of applying different categories of qualitative spatial relations in five aspects in three different tasks event classification, clustering, and detection has been compared. These five categories are: topology, direction, relative size, relative speed, and relative trajectories. According to their experimental results, the best performance is always achieved by a combination of all the five categories. Three categories of representations qualitative temporal, qualitative spatial, and quantitative spatial have been applied in [Tayyub et al., 2014] to recognized daily activities. Best recognition performance is achieved by combining all three categories.

Qualitative Action Recognition Explanation

Given a set of positive and negative examples (namely videos that are relevant or irrelevant to a specific action), (Dubba et al. 2015) has proposed an inductive logic programming framework that learns which object relation chains explain actions like *airplane arrival* or *airplane departure*. Or, in their terminology, a pipeline of methods is proposed to learn relational models for actions. Their solution involves manual annotation effort in labelling data, which can be problematic when dealing with large-scale data: *Deictic Supervision* requires a human to label roughly where and when actions occur in videos; a predefined object hierarchy is required; and some of the parameters (e.g. the weights of positive and negative examples) are decided manually. Their algorithm is also very sensitive to the initial positive example selected and is very likely to provide multiple predictions for a single video, which might lead to many false positive errors in action recognition.

In (Zhuo et al. 2019; Li et al. 2018), actions are qualitatively (or symbolically) described as state transitions. A state can be an object in a certain status (e.g. an open microwave) or a pair of object in a particular relationship (e.g. hand holding cup). The qualitative action descriptions are used as "prior knowledge" in action recognition and explanation. Specifically, first an object detector and state detector are trained; second, these two detectors are applied to obtain the state transitions of each object or object pair in each video; and third, an action is detected if a state transition satisfies the description of an action and this state transition is used as the explanation for the recognized action. They re-annotate the CAD120 dataset (Koppula, Gupta, and Saxena 2013) and experiment on multi-action-label videos from this new dataset. The experimental results show their approach can detect most actions (high recall) but also cause many false positive recognition errors (low precision).

Attention for Explanation

Attention is a mechanism (Vaswani et al. 2017) that has become popular and widely applied in deep learning and computer vision in recent years. Attention mechanism has been proved to be useful in improving the performance of recognition models (Selvaraju et al. 2017; Jetley et al. 2018;

Fukui et al. 2019). The intuition of attention is similar to human cognition in the sense that it helps recognition models find which parts of input data are more important and pay more "attention" on these deciding parts. Technically, by learning, an attention layer is able to assign greater weights to features that are critical in accurate recognition. By visualizing which features are more "attended", attention is able to improve the explainability of image classification neural networks (Selvaraju et al. 2017; Fukui et al. 2019).

Proposed Approach

As shown in Figure 4, our proposed approach consists of three main components: *input*, *saliency estimation*, and *outputs*. In each of the three components, there is an innovative module that plays a key role: *object relation chain extractor* in input, *neural saliency estimator* in saliency estimation, and *explanation generator* in outputs.

We divide each video into several clips (Bleser et al. 2015) and manipulate the object relation chains in each video clip separately. In this way, object relation chains in the same video clip are more likely to be semantically relevant; and it is possible to find salient object relation chains in different phases of an action.

The object relation chain extractor detects the set of object relation chains in each video clip. Then each set of extracted object relation chains is k-hot encoded into a vector so that it can be manipulated by a neural network. These k-hot encoded vectors constitute a matrix, which is the input of our neural saliency estimator. The output is a saliency score matrix whose values indicate which object relation chains are more salient in each video clip; and a saliency weighted object relation chain matrix so that the saliency scores of object relation chains can be utilized to boost performance in action recognition. With the saliency weighted object relation chain matrix as input, action in the video is recognized by several fully connected layers and their corresponding activation functions. The recognized action is justified by the explanations generated by our explanation generator.

Object Relation Chain Extraction

Assume we are given a set of qualitative spatial relations $\mathcal{R}$ that are defined in a certain set of qualitative spatial calculi, and a set of object categories $\mathcal{O}$. Each object relation chain c is a $2+n$ tuple o_1-o_2-r_1-r_2-...-r_n, where $o_1, o_2 \in \mathcal{O}$, $r_1, r_2, ..., r_n \in \mathcal{R}$, $n \in \mathbb{Z}^+$, and $r_i \neq r_j$ if $j = i+1$. It is actually a subset of the representation introduced in (Duckworth, Hogg, and Cohn 2019) that focuses on a consecutive set of changes in the relation between the same object pair.

The set of all possible object relation chains is definite given $\mathcal{O}, \mathcal{R}$, and n_r^*, where n_r^* is the maximum number of relations an object relation chain can have. An object relation extractor is a function $\mathcal{E}$ that maps a video clip into the set of object relation chains in it. I.e., $\mathcal{E}(\nu)$ denotes all object relation chains that can be observed in a video clip ν. It's not difficult to implement an object relation chain extractor since state-of-the-art computer vision algorithms (Wang et al. 2019; Li et al. 2019; Tan, Pang, and Le 2020) have already been able to detect and track object locations in a very accurate manner. Those object locations are mainly in the form

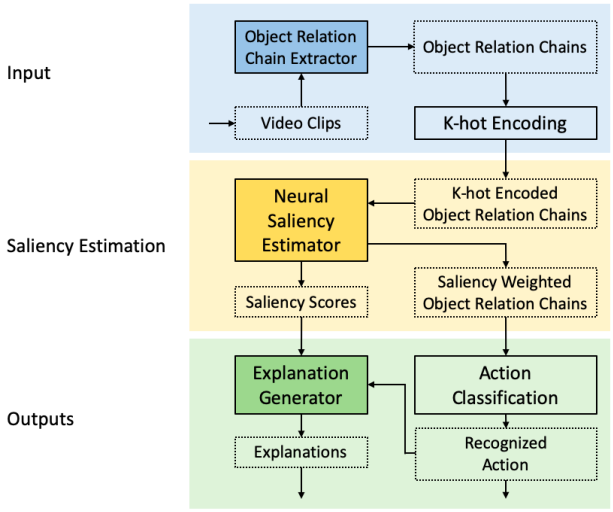

Figure 4: An overview of our approach: the input is a set of video clips and the outputs are a recognized action and its explanations. Our approach consists of three components: input, saliency estimation, and outputs. Solid rectangles are modules and dotted rectangles are data in different forms. Our innovative modules in each component are object relation chain extractor (blue), neural saliency estimator (yellow), and explanation generator (green).

of minimal bounding boxes, and qualitative spatial relations between objects are estimated by those between their corresponding bounding boxes. Relations between bounding boxes can be automatically calculated because most qualitative spatial relations have geometric definitions (Gatsoulis et al. 2016). Given the qualitative spatial relations in each frame of a video clip, object relation chains are extracted by looking for pairs of objects that are in different qualitative spatial relations in two consecutive frames.

Neural Encoding of Object Relation Chains

Given a video clip ν, we propose to represent its extracted object relation chains $\mathcal{E}(\nu)$ as a *k-hot* vector, $\boldsymbol{x} = [x_1, x_2, ..., x_N]^T$, where N is the total number of distinct object relation chains given $\mathcal{O}, \mathcal{R}$, and n_r^*; $x_i = 1$ iff the corresponding object relation chain is observed in ν, otherwise, $x_i = 0$. A video that is divided into n_ν clips $\{\nu_1, \nu_2, ..., \nu_{n_\nu}\}$ is represented as a $n_\nu \times N$ matrix $\boldsymbol{M}^v = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{n_\nu}]$, where $\boldsymbol{x}_i$ is the corresponding k-hot vector of ν_i. In this way, we compactly represent object relation chains in a video as a real-valued fixed-size matrix, which enables learning neural models on symbolic video representations.

Saliency Estimation and Action Recognition

As aforementioned, agents usually have a noisy observation of the target action in the presence of interfering or incorrect object relation chains. In order to achieve robust recognition in a noisy environment, we start from the observation that *the semantics of an object relation chain is context-dependent*. Namely, an agent has to jointly consider the co-occurrence of

multiple object relation chains, in order to understand actions correctly in context. One such example is shown in Figure 1: let o_1 be *car* and o_2 be *parking space*, if it occurs alone, the object relation chain o_1-o_2-*dc*-*po* is very likely caused by an accidental interaction (Figure 1b) while its co-occurrence with another object relation chain o_1-o_2-*po*-*pp* is a strong indicator of *car parking* (Figure 1a).

Following this observation, we propose an innovative *neural saliency estimator*, which measures the importance of observed object relation chains in action recognition, while taking into account other co-occurring object relation chains. It is actually a neural network layer that takes the encoding of observed object relation chains in a video clip as input and learns to re-weight them by considering exhaustively all the pair-wise co-occurrence of these chains.

As discussed above, the saliency of an object relation chain is relevant to all co-existing chains. This relevance is described in Equation 1, where $s_i \in \boldsymbol{s}$ is the saliency score of its corresponding object relation chain and a weighted sum of all co-existing chains. A higher M^s_{ij} indicates the occurrence of object relation chains i and j is highly correlated while lower values show such co-occurrence is likely accidental thus being noise for recognition.

$$s_i = \sum_{j=1}^{N} M^s_{ij} x_j \tag{1}$$

So, as shown in Figure 5, given N distinct object relation chains, the k-hot encoded object relation chain vector $\boldsymbol{x}$ is mapped into the saliency score vector $\boldsymbol{s}$ by multiplying with a $N \times N$ matrix $\boldsymbol{M}^{cs}$. Namely, $\boldsymbol{s} = \boldsymbol{M}^{cs}\boldsymbol{x}$. The matrix $\boldsymbol{M}^{cs}$ is named as the *pairwise co-occurrence saliency mapping matrix*. Then the saliency score vector $\boldsymbol{s}$ will be exponentially normalized into $\boldsymbol{s}'$ by the softmax function σ as in Equation (2), where $s_i \in \boldsymbol{s}$ and $s'_i \in \boldsymbol{s}'$.

$$s'_i = \sigma(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \tag{2}$$

Last, the k-hot encoded object relation chain vector $\boldsymbol{x}$ will be re-weighted by the normalized saliency scores. As in Equation (3), each element x^{sw}_i in the saliency weighted object relation chain vector $\boldsymbol{x}^{sw}$ is the result of its correspondence in the original object relation chain vector x_i multiplied by the corresponding normalized saliency score s'_i.

$$x^{sw}_i = s'_i x_i \tag{3}$$

Given a $n_\nu \times N$ object relation chain matrix $\boldsymbol{M}^v$, the outputs are a $n_\nu \times N$ normalized saliency score matrix $\boldsymbol{M}^{\sigma s}$ and a $n_\nu \times N$ saliency weighted object relation chain matrix $\boldsymbol{M}^{sw}$. As introduced earlier, the former will be used for explanation generation while the latter will be the input of action classification. We use the saliency scores to generate action explanations. For action recognition, we apply multiple fully connected layers to acquire the final action prediction.

Our neural-based saliency estimator and action classifier are trained together by minimizing the *cross entropy loss* in action classification, which is based on the assumption that a better saliency estimator leads to a more accurate action classifier. This assumption will be proved by experiments.

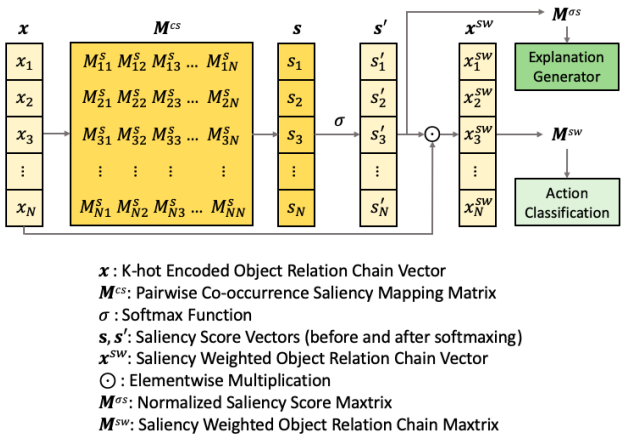

Figure 5: Our neural saliency estimator will be applied to each object relation chain vector in $\boldsymbol{M}^v$: each k-hot encoded $\boldsymbol{x}$ is mapped into a saliency score vector $\boldsymbol{s}$ by our pairwise co-occurrence saliency mapping matrix $\boldsymbol{M}^{cs}$. Then $\boldsymbol{s}$ is normalized by the softmax function. The normalized saliency score vector $\boldsymbol{s}'$ is used to constitute the normalized saliency score matrix $\boldsymbol{M}^{\sigma s}$ and produce the saliency weighted object relation chain vector $\boldsymbol{x}^{sw}$ by elementwise multiplied with $\boldsymbol{x}$. The saliency weighted object relation chain matrix $\boldsymbol{M}^{sw}$ consists of $\boldsymbol{x}^{sw}$ from each video clip and it is the input of our action classifier.

Explanation Generation

Our explanation generator produces action explanations in the form of salient object relation chains (i.e. object relation chains with the highest saliency scores). Using such an explanation generator, our action recognizer is explainable with salient object relation chains as our explanations. Such explanations are human-understandable. For example, they can be easily translated into natural language in the form of "this action is performed in the video because we believe this set of observed object relation chains is caused by it" and critically, the spatial relations composing these chains are intuitively understandable (Shariff, Egenhofer, and Mark 1998).

Given a video v, let $\boldsymbol{M}^{\sigma s}$ be the output of our neural saliency estimator and a be the output of the action classifier. Remember, each element in $\boldsymbol{M}^{\sigma s}$ corresponds to the saliency score of a certain object relation chain in a specific video clip. So with $\boldsymbol{M}^{\sigma s}$, we can construct a function $\mathcal{S}$ that maps an object relation chain c, a video clip ν, and an action a into a saliency score. For example, $\mathcal{S}(c, \nu, a) \in [0, 1]$ is a real number that indicates how salient c is if the action performed in v is predicted as a.

Our explanation for a certain action recognition is generated by the rule defined in (4), where ν is a video clip, and $\mathcal{S}_k(\mathcal{E}(\nu), \nu, a)$ denotes the kth highest saliency score among all the saliency scores of object relation chains in $\mathcal{E}(\nu)$. Namely, we will select the top-k salient object relation chains (i.e. object relation chains with top-k highest saliency scores) in $\mathcal{E}(\nu)$ as our explanation for a single video clip. The explanation for the whole video is an orderly combination of

the explanations for its video clips.

$$c \in \mathcal{X}(\nu, a) \Leftarrow \mathcal{S}(c, \nu, a) \geq \mathcal{S}_k(\mathcal{E}(\nu), \nu, a) \quad (4)$$

Our approach can also be applied to enhance the explainability of state-of-the-art action recognition models. If the prediction from our approach is the same as that from those approaches, the generated explanation describes the reasoning of the action recognition process. If not, their prediction is preferred and our explanation helps further examination.

Experiments and Evaluation

In this section, we conduct experiments on two datasets to evaluate our approach by comparing it with two other baseline explainable action recognition approaches.

Baseline Approaches. We compare our method with two baselines, *intuitive thresholding* and *intuitive maximum* that generate explanations for actions. They originate from the inductive and abductive reasoning algorithm in (Dubba et al. 2015) and we adapt it for our purposes as below.

For each action $a \in \mathcal{A}$, we refer to the set of videos labeled with a as its *positive examples* $P(a)$ and the set of videos not labeled with a as its *negative examples* $N(a)$. Formally, for each $a \in \mathcal{A}$, $P(a) = \{v | \mathcal{L}(v) = a, v \in \mathcal{V}\}$ and $N(a) = \{v | \mathcal{L}(v) \neq a, v \in \mathcal{V}\}$. These two approaches are based on the same *intuitive* assumption: the more frequent an object relation chain exists in the positive examples of an action and the less frequent it is in the negative examples of the same action, the more likely this object relation chain is caused by the action (Dubba et al. 2015). Based on this assumption, let $P(c, a) = \{v | c \in \mathcal{E}(v), v \in P(a)\}$ be the set of a's positive examples that contain c and $N(c, a) = \{v | c \in \mathcal{E}(v), v \in N(a)\}$ be the set of a's negative examples that contain c, the saliency score of an object relation chain c for an action a is denoted as $\mathcal{S}_{it}(c, a)$ and defined in Equation (5), where it is short for intuitive and η is a hyper-parameter to adjust the relative importance of these two ratios.

$$\mathcal{S}_{it}(c, a) = \frac{|P(c,a)|}{|P(a)|} - \eta \frac{|N(c,a)|}{|N(a)|} \quad (5)$$

These two approaches share the same saliency estimator as defined in (5) but are different in explanation generation and action recognition. *Intuitive thresholding* is based on the assumption that an object relation chain is caused by an action if the saliency score of an object relation chain for an action is higher than a certain threshold. Formally, let τ be the threshold and $\mathcal{D}(a)$ be the set of causal object relation chains for a, $\mathcal{D}(a)$ is constructed according to the assumption that $c \in \mathcal{D}(a) \Leftarrow \mathcal{S}_{it}(c, a) > \tau$. Given an unseen video v, the set of v's action labels $\mathcal{L}^{\tau}(v)$ is decided by the principle that $a \in \mathcal{L}^{\tau}(v)$ iff there exists c that satisfies both $c \in \mathcal{E}(v)$ and $c \in \mathcal{D}(a)$. Namely, a video v can be labelled with an action a if there is an object relation chain in v that is also in a's set of causal object relation chains $\mathcal{D}(a)$. An action explanation is a set of such causal object relation chains.

One obvious drawback of the first approach is: it is very likely an unseen video is labelled with multiple actions, which is a property inherited from the approach in (Dubba et al. 2015). To overcome this problem, similar to our approach, the second baseline approach *intuitive maximum* focuses on object relation chains with top 1 saliency scores: $\mathcal{L}(v) = a \Leftarrow c \in \mathcal{E}(v)$, $a \in \mathcal{A}$, and $\mathcal{S}_{it}(c, a) = max(\{\mathcal{S}_{it}(c_*, a_*) | c_* \in \mathcal{E}(v), a_* \in \mathcal{A}\})$. In this case, an action explanation consists of only one object relation chain (i.e. the one with the maximum saliency score and used to determine the recognized action). We report results of both approaches using optimized hyper-parameters for action recognition from grid search.

Datasets. We use two datasets in our experiments: CAD120 (Koppula, Gupta, and Saxena 2013) and CAD120++ (Zhuo et al. 2019). CAD120++ is a relabelled version of CAD120. CAD120 consists of 124 videos that record 10 actions (e.g. *making cereal* or *taking medicine*) performed by four different people. 12 categories of objects have been labelled including *hand*, *bowl*, *cup* and so on. In contrast, there are 551 videos and 10 different action labels in CAD120++. Actions (e.g. *pick* or *place*) in CAD120++ are much simpler than those in CAD120 in the sense that average number of frames per video is much smaller: 55.7 v.s. 525.3. Accordingly, videos in CAD120++ and CAD120 are divided into 1 and 10 video clips respectively. Only CAD120++ provides ground truth action explanations, which enables to evaluate the explanations generated by those three approaches.

In CAD120++, 7 out of 10 actions are explained in the form of changes in relations between objects: there is no explanation for *null* and the explanations for *open* and *close* rely on changes in object attributes (e.g. one of the given explanations for *open* is that the attribute of *microwave* has changed from *closed* to *open*). Ground truth action explanations in the form of object relation chains are generated by replacing natural language relations (e.g. *not holding*, *holding*, *apart*, or *contacting*) in CAD120++ explanations with RCC-4 relations. For example, one of the ground truth explanations of *pick* is *(hand,box,not holding,holding)*, which is translated into an object relation chain $hand$-box-dc-po. So, for each of the 7 actions, there is a ground truth explanation in the form of a set of object relation chains $\mathcal{X}_{gt}(a)$.

We choose RCC-4 as the ground truth qualitative spatial calculus because most of those natural language relations can be described by topological relations between objects and in a tentative experiment, the performance of RCC-4 is the best out of other RCC variations and other calculi like QDC-4 and CDR when it is individually used in action recognition.

Evaluation Metrics We follow the convention as in (Tayyub et al. 2014; Zhuo et al. 2019) to divide each dataset into 4 folds based on which person performs the action (as mentioned earlier, there are 4 different actors). Namely, we apply 4-fold cross validation and results are averaged across all the 4 folds. Action recognition is evaluated by the accuracy in classification. Saliency estimation accuracy (or the quality of generated explanations) is evaluated by *top k rate* ($k \in \{1, 2, 5\}$). Given a video v and its predicted action a, $\mathcal{X}(v, a)$ is a top-k explanation if there is at least one of its top-k object relation chains is in $\mathcal{X}_{gt}(a)$. Top k rate is the rate of recognized videos with top-k explanations out of all recognized videos.

Experimental Results and Analysis

A series of experiments have been conducted on CAD120++ and CAD120 and the results are listed in Table 1. The two

baseline approaches and our approach have been evaluated in Experiment 1, 2, and 3 respectively. One additional experiment has been conducted to further analyze the relationship between saliency estimation accuracy and action recognition accuracy and the results are shown in Figure 6.

Action Recognition Accuracy As in Table 1, action recognition accuracy is evaluated and our approach achieves much better performance on both datasets when compared to the two baseline approaches. This is mainly because the proposed neural saliency estimator captures the co-occurrence of object relation chains. In this way, our approach is less vulnerable to interfering or incorrect object relation chains thus achieving more robust and accurate recognition. These results also support that our approach is data efficient and works well on small-scale datasets like CAD120.

Quality of Explanation Generation As in Table 1, quality of generated explanations (or saliency estimation accuracy) is evaluated by "top1/2/5 rates". The top1/2/5 rates of our approach are much better than those of the two baseline approaches. Our explanation generator learns to model the relation between object relation chains and actions, which makes it more adaptive in diverse scenarios.

Table 1: Results of experiments on CAD120++ and CAD120. **No.**: experiment indexes. **approach**: the approach to action recognition and explanation generation. **intui_thres** is short for intuitive thresholding; **intui_max** is short for intuitive maximum; and **ours** stands for our approach. **accuracy** is the metric for action recognition while **top1/2/5 rates** is to evaluate generated explanations. Results in accuracy and top1/2/5 rates are in percentage (%) and best results are highlighted in bold.

No.	approach	CAD120++		CAD120
		top1/2/5 rates	accuracy	accuracy
1	intui_thres	53.6/57.6/67.2	53.9	67.3
2	intui_max	29.8/33.8/35.4	54.3	71.0
3	ours	**61.6/73.8/84.1**	**69.5**	**95.9**

Impact of Saliency Estimation on Recognition As shown in Figure 6, during the training process, action recognition accuracy and top1 rates are positively correlated. This indicates that recognition benefits from improved saliency estimation, which further supports the assumption that our neural saliency estimator plays a key role in action recognition.

Qualitative Analysis As shown in Figure 7, the ground truth object relation chain *microwave-cloth-po-ppi* is identified with the highest saliency score. As a result, the action *clean* in the top video is correctly recognized. As a failure case analysis, in the bottom video, our approach mistakenly estimates interactions between hand and microwave (i.e. the top two object relation chains) as salient object relation chains, which causes an incorrect prediction.

Conclusion and Future Work

In this paper, we have proposed an accurate neuro-symbolic approach that achieves explainable action recognition. Our main innovations are (1) object relation chain, a simple but

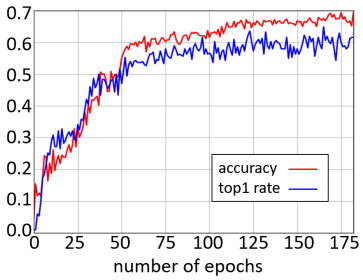

Figure 6: Changes in action recognition accuracy and top1 rates as the number of epochs increases. Results are from training our approach on CAD120++ with different number of epochs. The red and blue lines show the changes in action recognition accuracy and top1 rate respectively.

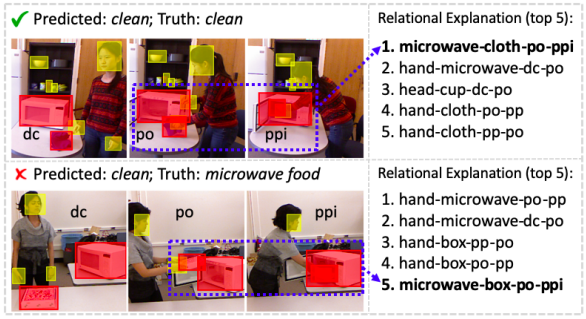

Figure 7: A correctly recognized action (top) and a wrongly recognized action (bottom). Right: top 5 salient object relation chains in the explanation. Ground truth object relation chains are in bold. Objects in red are involved in ground truth object relation chains while irrelevant objects are in yellow.

effective symbolic action video representation; (2) a neural saliency estimator that estimates the importance of object relation chains in action recognition; and (3) an explanation generator that generates action explanations in the form of salient object relation chains.

We believe that our approach makes an important contribution towards solving a more general problem: explainable classification by finding critical symbolic features. Our pipeline is to first extract symbolic features and then estimate their importance in classification, which can also be applied to many other similar problems that require trustworthy and explainable agents. Our approach is data efficient and performances well on accuracy and explainability in small-scale datasets. While in large-scale datasets our approach might not be as accurate as state-of-the-art learning approaches in action recognition, our approach can be used to augment these approaches and to provide them with explainabilities. In the future, we aim to make more contributions on this general problem by combining symbolic features with numeric features and utilizing large-scale knowledge databases.

References

Bleser, G.; Damen, D.; Behera, A.; Hendeby, G.; Mura, K.; Miezal, M.; Gee, A.; Petersen, N.; Maçães, G.; Domingues, H.; et al. 2015. Cognitive learning, monitoring and assistance of industrial workflows using egocentric sensor networks. *PloS one*, 10(6): e0127769.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, J.; Cohn, A. G.; Liu, D.; Wang, S.; Ouyang, J.; and Yu, Q. 2015. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(1): 106–136.

Clementini, E.; Di Felice, P.; and Hernández, D. 1997. Qualitative representation of positional information. *Artificial intelligence*, 95(2): 317–356.

Cohn, A. G.; and Renz, J. 2008. Qualitative Spatial Representation and Reasoning. In *Handbook of Knowledge Representation*, 551–596.

Cohn, A. G.; Renz, J.; and Sridhar, M. 2012. Thinking inside the box: A comprehensive spatial representation for video analysis. In *KR*.

Dubba, K. S.; Cohn, A. G.; Hogg, D. C.; Bhatt, M.; and Dylla, F. 2015. Learning relational event models from video. *Journal of Artificial Intelligence Research*, 53: 41–90.

Duckworth, P.; Hogg, D. C.; and Cohn, A. G. 2019. Unsupervised human activity analysis for intelligent mobile robots. *Artificial Intelligence*, 270: 67–92.

Dylla, F.; Lee, J. H.; Mossakowski, T.; Schneider, T.; Delden, A. V.; Ven, J. V. D.; and Wolter, D. 2017. A survey of qualitative spatial and temporal calculi: algebraic and computational properties. *ACM Computing Surveys (CSUR)*, 50(1): 7.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Frank, A. U. 1991. Qualitative spatial reasoning with cardinal directions. In *7. Österreichische Artificial-Intelligence-Tagung/Seventh Austrian Conference on Artificial Intelligence*, 157–167. Springer.

Freksa, C. 1992. Using orientation information for qualitative spatial reasoning. In *Theories and methods of spatio-temporal reasoning in geographic space*, 162–178. Springer.

Fukui, H.; Hirakawa, T.; Yamashita, T.; and Fujiyoshi, H. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10705–10714.

Galata, A.; Cohn, A. G.; Magee, D. R.; and Hogg, D. C. 2002. Modeling Interaction Using Learnt Qualitative Spatio-Temporal Relations and Variable Length Markov Models. In van Harmelen, F., ed., *Proceedings of the 15th Eureopean Conference on Artificial Intelligence, ECAI'2002, Lyon, France, July 2002*, 741–745. IOS Press.

Gatsoulis, Y.; Alomari, M.; Burbridge, C.; Dondrup, C.; Duckworth, P.; Lightbody, P.; Hanheide, M.; Hawes, N.; Hogg, D.; Cohn, A.; et al. 2016. QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video.

Jetley, S.; Lord, N. A.; Lee, N.; and Torr, P. H. 2018. Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

Koppula, H. S.; Gupta, R.; and Saxena, A. 2013. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8): 951–970.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 4282–4291.

Li, D.; Rodriguez, C.; Yu, X.; and Li, H. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1459–1469.

Li, D.; Scala, E.; Haslum, P.; and Bogomolov, S. 2018. Effect-abstraction based relaxation for linear numeric planning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4787–4793.

Li, D.; Xu, C.; Yu, X.; Zhang, K.; Swift, B.; Suominen, H.; and Li, H. 2020b. TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation. In *Advances in Neural Information Processing Systems*, volume 33.

Li, D.; Yu, X.; Xu, C.; Petersson, L.; and Li, H. 2020c. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6205–6214.

Li, R.; Hua, H.; Haslum, P.; and Renz, J. 2021. Unsupervised Novelty Characterization in Physical Environments Using Qualitative Spatial Relations. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021*, 454–464.

Ligozat, G. 1998. Reasoning about Cardinal Directions. *J. Vis. Lang. Comput.*, 9(1): 23–44.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 7083–7093.

Moratz, R.; and Ragni, M. 2008. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1): 75–98.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6964–6974.

Randell, D. A.; Cui, Z.; and Cohn, A. G. 1992. A spatial logic based on regions and connection. *KR*, 92: 165–176.

Renz, J.; and Mitra, D. 2004. Qualitative direction calculi with arbitrary granularity. In *PRICAI*, volume 3157, 65–74.

Renz, J.; and Nebel, B. 2007. Qualitative Spatial Reasoning Using Constraint Calculi. In *Handbook of Spatial Logics*, 161–215.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shariff, A. R. B.; Egenhofer, M. J.; and Mark, D. M. 1998. Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms. *International journal of geographical information science*, 12(3): 215–245.

Skiadopoulos, S.; and Koubarakis, M. 2004. Composing cardinal direction relations. *Artificial Intelligence*, 152(2): 143–171.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2010. Unsupervised learning of event classes from video. In *AAAI*.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *CVPR*, 10781–10790.

Tayyub, J.; Tavanai, A.; Gatsoulis, Y.; Cohn, A. G.; and Hogg, D. C. 2014. Qualitative and quantitative spatio-temporal relations in daily living activity recognition. In *ACCV*, 115–130. Springer.

Tran, D.; Wang, H.; Torresani, L.; and Feiszli, M. 2019. Video classification with channel-separated convolutional networks. In *ICCV*, 5552–5561.

Van Beek, P. 1992. Reasoning about qualitative temporal information. *Artificial intelligence*, 58(1-3): 297–326.

Van de Weghe, N.; Cohn, A.; De Tre, G.; and De Maeyer, P. 2006. A qualitative trajectory calculus as a basis for representing moving objects in geographical information systems. *Control and Cybernetics*, 35(1): 97–119.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; and Torr, P. H. 2019. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 1328–1338.

Young, J.; and Hawes, N. 2015. Learning by observation using qualitative spatial relations. In *AAMAS*, 745–751. AAMAS.

Zhuo, T.; Cheng, Z.; Zhang, P.; Wong, Y.; and Kankanhalli, M. 2019. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th ACM International Conference on Multimedia*, 521–529.