Action Recognition from Skeleton Data Via Analogical Generalization

Kezhen Chen and Kenneth D. Forbus

Qualitative Reasoning Group, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208 USA

Abstract

Human action recognition remains a difficult problem for AI. Traditional machine learning techniques have had some success, but have two disadvantages. First, these models are typically black boxes whose internal models are not inspectable and whose results are not explainable. Second, typically massive amounts of data are needed to achieve good recognition performance. This paper describes a new pipeline for recognizing human actions from skeleton data via analogical generalization. Specifically, starting with Kinect data, we segment each human action by finding temporal regions where the qualitative spatial descriptions are constant, creating a *sketch graph* that provides a compact relational representation of the behavior that is easy to visualize. Models are learned from sketch graphs via analogical generalization, which are then used for classification via analogical retrieval. The retrieval process also produces links between the new example and components of the model that provide explanations. We describe results on two public standard datasets to illustrate its utility.

Introduction

Human action recognition is an important but difficult problem. Traditional machine learning techniques rely on extracting large numbers of features and using techniques such as deep learning [Baccouche *et al*., 2011]. However, these techniques have two disadvantages. First, they are black boxes: They can produce results, but they provide no explanations for their answers. This makes their results difficult to trust and to debug [Lowd *et al*., 2005]. Second, they typically require massive amounts of data to achieve reasonable accuracy. Such data is not always available. People learn with far less data than today's machine learning systems require [Forbus *et al.*, in press]. We suspect that part of the reason is that, even for visual tasks, relational representations are important in human cognition [Marr, 1982; Palmer 1999]. By working with relational representations inspired by human vision, can we produce more explainable results?

This paper draws on research in qualitative spatial reasoning and cognitive simulation of visual problem-solving and analogy to provide a new approach to recognizing human actions from Kinect skeleton data. Instead of computing frame-based features, the video stream is divided into a *sketch graph*, consisting of multiple sequences of snapshots. Each snapshot is like a panel in a comic strip: It consists of a motion described by a single qualitative state, which might correspond to one frame or many. Each body segment has its own sequence of such states. The trajectories within these states and relationships across these states are described qualitatively, using automatically constructed visual representations. The sketch graphs for each instance of a behavior type are combined via *analogical generalization*, to automatically construct probabilistic relational schemas (plus outliers) characterizing that behavior type. Given a new behavior, a sketch graph is computed for it, and analogical retrieval is used across the entire set of behavior models to retrieve the closest schema (or outlier). The correspondences found during analogical retrieval provide links between the new example and the aspect of the model used, providing a means of explanation.

We begin by summarizing the work we build on, including the Qualitative Trajectory Calculus, CogSketch, and analogical processing. We then describe the learning pipeline and how classification works. Results on the University of Texas at Dallas(UTD) Multimodal Human Action and the Game 3D datasets are described, and we close with related and future work.

Background

Our approach combines ideas from qualitative spatial reasoning, visual problem solving, and analogical processing. We discuss each in turn.

The Qualitative Trajectory Calculus (QTC)

QTC [Weghe *et al*., 2005] is a qualitative calculus for representing the motion of a continuously moving point object in a Euclidean space relative to a stationary reference object.

Given a moving object W and a reference point P, at any time the relative motion of W can be described as either:

- -: W is moving towards P
- +: W is moving away from P
- 0: W is stable with respect to P

In QTC theory for three dimensions, moving towards relation means that the Euclidean distance between the object and the reference decreases, moving away relation means that the Euclidean distance between the object and the reference increases and stable relation means that the Euclidean distance between the object and the reference does not change. Consequently, if W is stable with respect to P and P has 0 velocity, W could also be stationary object with 0 velocity or W could move around P with a circle path. In both situations, the Euclidean distance between W and reference P does not change.

CogSketch

CogSketch [Forbus et al., 2011] is a sketch understanding system that provides a model of high-level visual processing. It provides multiple, hierarchical levels of visual representation, including decomposing digital ink into edges, combining edges into entities, and gestalt grouping methods. The qualitative visual representations that it automatically computes from digital ink have enabled it to model a variety of visual problem-solving tasks [e.g. Lovett & Forbus, 2011; 2017]. These relations include qualitative topology [Cohn et al., 1997], positional relations (e.g. above, leftOf), and directional information (e.g. quadrants and directions).

Sketches in CogSketch can be divided into units, called *subsketches*, which themselves can participate in relationships. Here subsketches are used to implement the panels in the sketch graph, with digital ink in each subsketch corresponding to trajectory information. Thus, CogSketch's visual processing is used to construct additional relations among and between subsketches. The *metalayer* in CogSketch enables multiple subsketches and relationships between them to be displayed, to support visualization.

Analogical Processing

We build on models inspired by Gentner's [1983] structure-mapping theory of analogy and similarity. Its notion of comparison is based on structured descriptions, including both attributes and relations. There is considerable psychological evidence supporting it. This makes structure-mapping attractive for use in AI systems so that, with the right representations, what looks similar to us will look similar to our software and vice-versa. We use the Structure-Mapping Engine [SME; Forbus et al., 2016] for analogical matching, MAC/FAC [Forbus et al., 1995] for analogical retrieval, and SAGE [McLure et al., 2015] for analogical generalization. Since these operations are at the heart of our learning approach, we summarize each in turn.

SME takes as input two structured, relational representations and produces one or more *mappings* that describe how they align. These mappings include correspondences, i.e. what goes with what, a similarity score, and candidate inferences that suggest how statements from one description can be projected to the other. SME has been used in a variety of AI systems and cognitive models. Most relevant to this paper, the representations produced by CogSketch have been used to model human performance on several visual tasks, including Ravens' Progressive Matrices, one of the most common tests used to measure human fluid intelligence. The CogSketch model uses SME at multiple levels of visual representations, including re-representing visual descriptions automatically as needed, leading to performance in the 75% percentile, better than most adult Americans [Lovett & Forbus, 2017]. Ravens and the other visual problems that SME has been used with are static, this paper marks the first time it and the other analogical components have been used with dynamic visual data.

Analogical retrieval is performed by MAC/FAC, which stands for "Many are Called/Few are Chosen", because it uses two stages of map/reduce for scalability. The inputs consist of a probe case and a case library. The MAC stage computes, in parallel, dot products over vectors that are automatically constructed from structured descriptions, such that each predicate, attribute, and logical function are dimensions in the vector and whose magnitude in each dimension reflects their relative prevalence in the original structured description. The best, and up to two others (if they are sufficiently close) are passed to the FAC stage. The map component compares the best structured descriptions from the MAC stage to the input probe using SME. Again, the best match, with up to two others if sufficiently close, are returned. This provides scalability (because the MAC stage is inexpensive) as well as structural sensitivity (because the content vector dot product is a coarse estimate of SME similarity, followed by using SME itself).

Analogical generalization is performed by the Sequential Analogical Generalization Engine [SAGE; McLure et al., 2015]. Every concept to be learned by analogy is represented by a *generalization pool*, which maintains both generalizations and outlying examples. Examples are added incrementally. The closest matching item (example or generalization) is retrieved via MAC/FAC, using the contents of the generalization pool as a case library. If there is no item, or the similarity to what is retrieved is less than an *assimilation threshold*, the new example is added as an outlier. Otherwise, if the item retrieved is an example, the two are combined into a new generalization. This process involves merging them, replacing non-identical entities by skolems, and assigning a probability to each statement depending on whether it was in just one description or both. If the item

retrieved was a generalization, that generalization is updated with skolems and probabilities based on its alignment with the new example. Generalizations in SAGE are thus probabilistic, but still concrete – skolem entities may become more abstract due to fewer high-probability statements about them, but logical variables are not introduced. Instead, candidate inferences are used for schema application.

SAGE also supports classification, by treating the union of generalization pools as a large case library. The case library which contained the closest item is taken as the classification of that example, with the correspondences of the match constituting an explanation of why it is a good match. Since a generalization pool can have multiple generalizations, SAGE naturally handles disjunctive concepts.

Our Approach

Our approach focuses on human skeleton action recognition by using analogical generalization over qualitative representations. It is implemented as a pipeline with three stages: *Action Segmentation*, *Relational Enrichment*, and *Action Generalization*. The system is totally automatic, and all sketches and relations are computed from Cogsketch automatically. Figure 1 shows the pipeline of our system. We describe each stage, and classification, in turn.

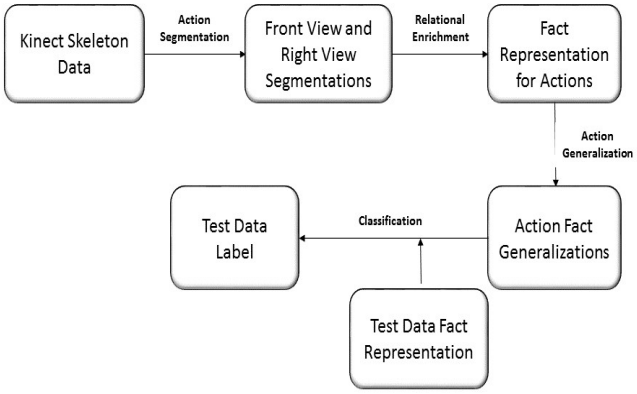

Figure 1: Flowchart for our pipeline system

Action Segmentation

The skeleton data produced by a Kinect (or other 3D sensors) contains many points per frame, representing each body part such as the head, neck, right hand, and right foot. We use 20 main body points to track movement of 20 body parts, as shown in Table 1. Connecting these body points provide a concise body skeleton graph. Each instance of an action consists of a continuous movement stream, sampled via many frames, each frame containing coordinates for these points. The first step of our pipeline abstracts away from frames into qualitatively distinct intervals describing the motion of particular body parts. A *track* is a sequence of data about a body part. For example, we use movements of right hand, left hand, right foot and left foot with respect to a static reference point as tracks. The quantitative description of motion represented by the change in track values frame by frame is used to construct qualitative descriptions of the motion of each track. Intervals of time over which the motion is qualitatively the same are *panels* in the sketch graph.

1	Head	6	Elbow Left	11	Wrist Right	16	Foot Left
2	Neck	7	Wrist Left	12	Hand Right	17	Hip Right
3	Spine Base	8	Hand Left	13	Hip Left	18	Knee Right
4	Spine Mid	9	Shoulder Right	14	Knee Left	19	Ankle Right
5	Shoulder Left	10	Elbow Right	15	Ankle Left	20	Foot Right

Table 1: Kinect 20 body points

We use head, spine-middle and spine-base as three reference points for computing QTC relations, using QSRlib [Gatsoulis et al., 2016]. For example, in segmenting the movement of the right hand, the head is picked as a reference point. The QTC relations can be computed for the right hand with respect to the head in each frame as following:

[0,0,0,+,+,+,+,0,0,0,0,-,-]

This track has 13 frames. "0" means that right hand is relatively stable with respect to head. "+" means that right hand moves towards the head and "–" means that right hand moves away from the head. In this example, the right-hand movement is segmented into four sub-actions. The right hand is static in first three frames, it moves towards the head in the next four frames, then it stay the same distances with respect to head in the next four frames and the right hand moves away from the head in the last two frames. The start and end times for each panel are recorded. Note that tracks will often have segmentations that vary from each other.

The next stage uses CogSketch to construct additional representations, but CogSketch assumes its input consists of 2D information. To provide an approximation to 3D, our system segments the movement in each skeleton data from front view based on (x,z) and from right view based on (z,y). It is also useful to simplify the segmentations, eliminating small segments that are very likely to represent noise, by computing the spatial distance between the start and end points for the segmentations for each track. The

segmentation with the highest variance is used to set a threshold for filtering small segments. Figure 2 shows a segmentation of raising right hands movement. Figure 2 (a) is the segmentation of front view and 2(b) is the segmentation of right view.

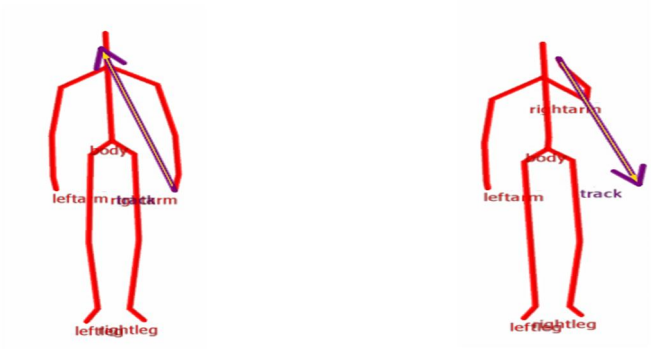

Figure 2 (a): front segmentation of raising hand(Kinect mirror view)

Figure 2 (b): right segmentation of raising hand(Kinect mirror view)

Relational Enrichment

The relational enrichment stage involves automatically adding additional relationships, using CogSketch, to provide more information about the motions within and between segments. Each example of a behavior is imported into CogSketch as a set of sketches, one per track, with each panel within a track being represented by a separate subsketch. Hence each action is represented by eight sketches in CogSketch: right hand front view, right hand right view, left hand front view, left hand right view, right foot front view, right foot right view, left foot front view and left foot right view. In each sketch and subsketch, CogSketch is used to compute relationships between body parts, e.g. the relative position of the right hand to the head.

We use the following logical function to denote panels in a sketch graph:

(KinectMotionFn <*body-part*> <*view*> <*direction*> <*token*>)

<*body-part*> is from the four main body points: right hand, left hand, right foot and left foot. <*view*> is front or right. <*direction*> is drawn from the CogSketch 2D direction vocabulary, which contains quadrant representations Quad1, Quad2, Quad3, Quad4, and Up, Down, Left and Right, plus the constant **NoMotion** to indicate there is no motion. <*token*> is a unique identifier denoting the segment.

Sequence information between panels is represented using the **occursAfter** relation, e.g.

(occursAfter
(KinectMotionFn RightHand Front Up D1RHFSeg2)
(KinectMotionFn RightHand Front Down D1RHFSeg4))

That is, a down movement of the right hand in this behavior occurs after an up movement. Additional details can be provided via Cyc's holdsIn relation[1]. For example, the spatial relations are helpful to decide the accurate locations of body points, so these relations can be added in the specific motion segments. The fact can be written as following:

(holdsIn
(KinectMotionFn RightHand Front Up D1RHFSeg2)
(Above Head RightHand))

This representation enables facts about different segments to be included in one unified case representation, without contradiction.

The attributes and relations that should be included for characterizing motion are still in flux, but our initial choices have been useful enough to get reasonable results, as illustrated below.

Action Generalization

All facts for each segment are combined as a case representing the entire action. Each such action is added to the generalization pool being used to learn that concept. For the experiments reported here, we used an assimilation threshold of 0.8. SAGE also uses a probability cutoff, i.e. if a fact's probability drops below this value, it is removed from the generalization. In these experiments, we used a probability cutoff of 0.2. Figure 3 shows a generalization pool which contains one generalization created from eight training examples. The other eight examples are outliers, sufficiently different from the generalization that they are stored as cases in the generalization pool.

[1] CogSketch uses the OpenCyc ontology and knowledge base contents, but a different reasoning engine.

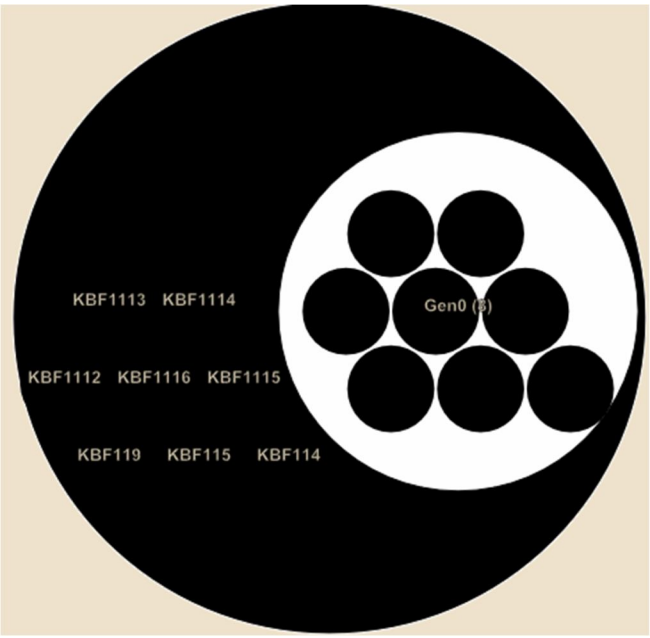

Figure 3: One example of a generalization pool

Classification

By treating the union of generalization pools for actions as one large case library, our system can classify new examples based on which library the closest retrieved item came from, as outlined above. The correspondences of the mapping constructed during analogical retrieval can be used to generate an explanation of why this label was deemed appropriate. We note that this explanation only includes positive information; an extension to SAGE enables it to construct near-misses [McLure et al., 2015], but we have not tried that extension on this task yet.

Experiments

While we view the ability to produce understandable explanations as an important part of our approach, we note that other approaches do not explore explanation, so we confine ourselves here to comparing using their metrics.

UTD-MHAD Dataset

The University of Texas at Dallas(UTD) Multimodal Human Action Dataset (UTD-MHAD) was collected as part of research on human action recognition by Chen et al. [Chen et al., 2015], University of Texas at Dallas. This dataset contains eight different humans performing twenty-seven actions in a controlled environment and each action repeats four times, collecting data from a Kinect V1 sensor. As our research only focuses on Skeleton data, we only use Skeleton files from RGB, Depth and Skeleton videos. There are many pairs of actions that may have same segments and similar representations, such as Wave and Throw, arm cross and Arm curl, and Draw circle clockwise and counter clockwise. To test the performance of our system training model, we pick the top seven significant and consistent motions in their task: swipe right, draw triangle, bowling, boxing, arm curl, catch and stand to sit.

The qualitative spatial relations and RCC8 relations are computed for the start point and end point of every movement track with respect to the Head, Spine-Mid and Spine-Base. The RCC8 relations are picked for the movement body point with respect to the other three movement body points, such as the RCC8 relations between right hand and left hand.

With the qualitative spatial relations described previously, we used the same cross-subject testing method from Chen, et al's [2015] paper, which contains four subjects' data for training and four subjects' data for testing. With the seven actions they used, we get 81% accuracy. We also used three subjects and two subjects as training data to test whether our system could have good performance even with less training data. We compared our results with the results of the same seven actions using a CRC classifier from the original paper, as shown in Table 2:

Methods	Training data number	Testing data number	Actions number	Overall Accuracy
Kinect & Inertial	4 subjects (16 data)	4 subjects (16 data)	7 actions	80.9%
Our method	4 subjects (16 data)	4 subjects (16 data)	7 actions	81%
Our method	3 subjects (12 data)	5 subjects (20 data)	7 actions	65.7%
Our method	2 subjects (8 data)	6 subjects (24 data)	7 actions	63.7%

Table 2: Results on UTD-MHAD dataset

Game 3D Dataset (G3D)

The G3D dataset contains 10 subjects performing 20 gaming actions collected with Microsoft Kinect. The 20 gaming actions are categorized into seven scenarios: fighting, golf, tennis, bowling, first person shooter, driving a car, miscellaneous. This dataset was collected as part of Bloom et al.'s [2014, 2016] online actions recognition research. They used five actions from fighting scenarios as their data to test their Hierarchical transfer learning algorithm.

We used the same pipeline for this data set, and used half of data as training data and the other half data for testing, focusing on the same five actions as Bloom et al.'s work: right punch, left punch, right kick, left kick and defend. We also trained our system with 1/3 of the data to test whether

our system still has reasonable performance. The results are presented in the following Table 3.

Methods	Training data	Actions number	F1 Score
Frame based	5 subjects	5 actions (fighting)	70.46%
Dynamic Feature Selection	5 subjects	5 actions (fighting)	91.9%
Our Method	5 subjects	5 actions (fighting)	77.6%
Our Method	3 subjects	5 actions (fighting)	71.58%

Table 3: Results on G3D dataset

Our method does better than their frame-based technique, but not as well as their dynamic feature selection model. On the other hand, with only three subjects' worth of data, it is already performing reasonably well. We suspect that using a more dynamic approach in the relational enrichment stage of our pipeline could boost performance further, but that is an empirical question.

Discussion

Our system has three advantages over traditional action recognition systems. First, given similar amounts of training data, it can achieve performance that is competitive with state of the art for several models. Moreover, it starts producing reasonable results with much less data. The second advantage is explanation. The constituents of the analogical model are relational representations, which can be displayed as text or via CogSketch, as opposed to a set of opaque numerical values for a classifier's parameters. Such explanations offer the possibility of enabling its users (or ultimately even itself) to tune its representations to improve performance. The third advantage is incremental operation: SAGE is designed for incremental operation, so that it can take advantage of new training examples without starting over from scratch.

Our approach requires reasonably consistent and accurate segmentation. If the initial data is too noisy, it will lead to many segments, among which many relationships might hold, which would make Action Segmentation and Relational Enrichment, and Action Generalization all slower and noisier. How well this approach scales to noisier data is an empirical question. We also plan to explore dynamic strategies for relational enrichment, using CogSketch's range of representations and data from the generalization pools to tune encoding strategies for better learning.

Related Work

Our paper shows a new method on the problem of human action recognition from a sequence of skeleton frame data, especially Kinect skeleton data. Actually, human action recognition from Kinect Skeleton data is a very popular topic and various methods have been used on this problem. In [Pichao Wang et al., 2016], the spatial-temporal information from 3D skeleton data was projected into three 2D images (Joint Trajectory Maps), and Convolutional Neural Networks were used for action recognition system. In [Ye, 2016], a novel octree-based algorithm was explored to extract spatial and temporal relations from each sequential Kinect 3D data, and the First-Take-All(FTA) feature vector was built for activity recognition with nearest neighbor search based on Hamming distance. In [Jiawei Li et al., 2016], the authors proposed a method on human action recognition by clustering human joints from Kinect data and use machine learning on each cluster to build a recognition system. Many researchers have focused on what kinds of features should be extracted in Skeleton movement data recognition. In [Maldonado et al., 2016], a feature selection algorithm, Reduction of Feature Dimensions based on Standard Deviation, was proposed to help extract useful features on human action recognition tasks. To the best of our knowledge, this is the first paper to do skeleton movement recognition via analogical reasoning instead of machine learning.

We used QSRlib to segment skeleton movements based on the change of qualitative relations. The idea of object and human movement analysis via qualitative relations has previously been used in many papers. Duckworth et al. [2016] describes an experiment that deployed a mobile robot in an office environment for 6 weeks to test many different qualitative representations and study which qualitative representation is best for learning human motion behaviors. [Thippur et al., 2015] and [Kunze et al., 2014] designed experiments to test representations for classifying environments from visual data via QSRlib. In [Dondrup et al., 2015], QSRlib was used to compute qualitative relations for a robot in its navigation system in order to generate a safe path between the robot and moving human. However, these prior efforts only used QSRLib for qualitative representations. Here, we use QSRlib to segment movements, but also use CogSketch to compute additional relationships within and across segments to provide richer relational description of the data, which facilitates learning.

Conclusion and Future Work

This paper presents a new approach, based on qualitative representations and analogical generalization, for learning how to classify human actions. Our three-stage pipeline

uses prior advances in qualitative spatial reasoning to segment tracks, a cognitive model of human high-level vision to enrich descriptions of motion and configuration, and analogical generalization to provide learning via inspectable, relational models. Our experiments provide evidence for the utility of this approach.

There are several avenues to explore next. The first is to test it with additional datasets, both to explore the noise and dynamic encoding issues. The second is to construct visualizations based on sketch graphs, to explore how they might be used to explain a system's classification to users. Furthermore, we plan to explore using this same approach to analyze video more broadly, including RGB and depth data. Finally, to explore tinier human movements, we will explore new models to represent human face movements and other specific movements.

Acknowledgements

This research was supported by the Machine Learning, Reasoning, and Intelligence Program of the Office of Naval Research.

References

[Alomari et al., 2016] Y Gatsoulis, M Alomari, C Burbridge, C Doudrup, P Duckworth, P Lightbody, M Hanheide, N Hawes, D C Hogg, A G Cohn. (2016). QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video. *QR.*

[Baccouche et al., 2011] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, Atilla Baskurt. (2011). Sequential Deep Learning for Human Action Recognition. In *Human Behavior Understanding* (pp. 29-39).

[Bloom et al., 2012] Victoria Bloom, Dimitrios Makris, Vasileios Argyriou. (2012). G3D: A gaming action dataset and real time action recognition evaluation framework. *CVPR.*

[Bloom et al., 2014] Victoria Bloom, Dimitrios Makris, Vasileios Argyriou. (2014). Clustered Spatio-Temporal Manifolds for Online Action Recognition. *ICPR*, (pp. 3963-3968).

[Bloom et al., 2016] Victoria Bloom, Vasileios Argyriou, Dimitrios Makriss. (2016). Hierarchical transfer learning for online recognition of compound actions. *Computer Vision and Image Understanding*, 62-72.

[Cohn et al., 1997] Anthony G. Cohn, Brandon Bennett, John Gooday, Micholas Mark Gotts. (1997). Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. In *GeoInformatica* (pp. 275-316).

[Chen et al., 2015] Chen Chen, Roozbeh Jafari, Nasser Kehtarnavaz. (2015). UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. *IEEE International Conference on Image Processing.*

[David Marr, 1982] David Marr. (1982). *Vision*. Freeman.

[Duckworth et al., 2016] Paul Duckworth, Yiannis Gatsoulis, Ferdian Jovan, Nick Hawes, David C Hogg, Anthony G Cohn. (2016). Unsuperivised Learning of Qualitative Motion Behaviours by a Mobile Robot. *AAMAS.*

[Forbus et al., 1995] Kenneth D. Forbus, Dedre Gentner, Keith Law. (1995). MAC/FAC: A Model of Similarity-based Retreval. *Cognitive Science*, 141-205.

[Forbus et al., 2016] Kenneth D. Forbus, Ronald W. Ferguson, Andrew Lovett, Dedre Gentner. (2016). Extending SME to Handle Large-Scalle Cognitive Modeling. *Cognitive Science*, 1-50.

[Forbus et al., 2011] Kenneth Forbus, Jeffrey Usher, Andrew Lovett, Kate Lockwood, Jon Wetzel. (2011). CogSketch: Sketch Understanding for Cognitive Science Research and for Education. *Topics in Cognitive Science*, 648-666.

[Forbus et al., in press] Kenneth Forbus, Chen Liang, Irina Rabkina. (in press) Representation and Computation in Cognitive Models. *Topics in Cognitive Science.*

[Gentner, 1983] Dedre Gentner. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

[Kunze et al., 2014] Lars Kunze, Chris Burbridge, Marina Alberti, Akshaya Tippur, John Folkesson, Patric Jensfelt, Nick Hawes. (2014). Combining Top-down Spatial Reasoning and Bottom-up Object Class Recognition for Scene Understanding. *IROS.*

[Lowd et al., 2005] Daniel Lowd; Christopher Meek. (2005). Adversarial Learning. *ACM.* Chicago.

[Lovett et al., 2017] Andrew Lovett, Kenneth Forbus. (2017). Modeling Visual Problem-Solving as Analogical Reasoning. *Psychological Review.*

[Li, 2016] Jiawei Li, Jianxin Chen. (2016). Joint Motion Similarity (JMS)-Based Human Action Recognition Using Kinect. *DICTA.*

[McLure et al., 2015] Matthew D. McLure, Scott E. Friedman, Kenneth D. Forbus. (2015). Extending Analogical Generalization with Near-Misses. *AAAI.* Austin: Texas.

[Maldonado et al., 2016] Carolina Maldonado, Homero Vladimir Rios-Figueroa, Antonio Marin-Hernandez. (2016). Improving action recognition by selection of features. *ROPEC.*

[Palmer et al., 1999] Stephen Palmer. (1999). *Vision Science: Photons to Phenomenology*. MIT Press.

[Thippur et al., 2015] Akshaya Thippur, Chris Burbridge, Lars Kunze, Marina Alberti, John Folkesson, Patric Jensfelt, Nick Hawes. (2015). A Comparison of Qualitative and Metric Spatial Relation Models for Scene Understanding. *AAAI.*

[Wang et al., 2016] Pichao Wang, Wangqing Li, Chuankun Li, Yonghong Hou. (2016). Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks. *IEEE Transactions on Cybernetics.*

[Weghe et al., 2005] N. Van de Weghe, A. Cohn, B. De Tre, P. De Maeyer. (2005). A Qualitative Trajectory Calculus as a basis for representing moving objects in Geographical Information Systems. *Control and Cybernetics*, 97-120.

[Ye, 2016] Jun Ye. (2016). Spatial and Temporal Modeling for Human Activity Recognition from multimodal Sequential Data.