SMALL HUMAN GROUP DETECTION AND EVENT REPRESENTATION BASED ON COGNITIVE SEMANTICS

Yafeng Yin, Guang Yang, Hong Man

ECE Department, Stevens Institute of Technology, Hoboken, NJ, 07030

ABSTRACT

We proposed a novel video event representation based on cognitive semantics for small human group detection and event recognition in this paper. Instead of video event symbols, the video is described via the basic cognitive elements, such as paths, places, things, actions, and causes. The structural and semantic distance of each thing in the same place will be calculated and different things (human in this case) will be merged together to reduce the semantic disorder. Once a human group is detected, its actions will be classified into atom group activities by their corresponding spatial and temporal semantics. The spatial and temporal similarity of atom group activities is examined and probabilistic context free grammar is derived from these atom activities based on Minimum Description Length (MDL) criterion. The induced grammar rules will then be used to parse test videos. The experimental results on the BEHAVE and Collective data set demonstrate the effectiveness of the proposed method.

Index Terms— Cognitive Linguistic Semantics, Context Free Grammar, Grammar Induction, Small Human Group Behavior

1. INTRODUCTION

Grammar models have attracted much attention in recent years for complex visual event recognition. To apply grammar models for event recognition, usually low-level features are firstly extracted from videos and then classified to a set of terminal symbols, i.e. visual event primitives. Different event primitives will form a discrete symbol string for syntactic analysis, including grammar induction and parsing [1]. This approach has been extended to more complex event recognition applications, such as the towers of Hanoi task [2], one-to-one basketball [3], office daily activities [4], and Blackjack games[5]. In summary, this approach is most suitable for sequential event recognition, regardless of the number of anticipants of the event. However, there are many scenarios with concurrent events, such as a small group of people fighting each other, namely "group fighting". The sub-event for each person cannot be treated separately and sequentially. So the simple sequential approach has difficulties to address such problems.

To recognize parallel visual events, Joo et.al. [6] introduced attribute grammar for event recognition and anomaly detection. Recently, Zhang.et.al [7] extended SCFG to automatically learning of grammar rules and parallel parsing of sub-events simultaneously. Besides the temporal semantics, spatial semantics have also been introduced to recognize two-person interactions [8]. These methods added attributes to each event primitive, for example, an ID set is stored in [7] and used for searching other concurrent event during parsing process. These approaches are suitable for a small number of parallel sub-events, and are very specific to certain applications.

Inspired by the cognitive linguistic (CL) models [9, 10], we introduce a CL based representation for visual events in videos. Five different conceptual primitives, including place, path, action, thing and cause, are used to represent different visual events as shown in Figure 1. As this representation is learned from the most fundamental constructor of human language, it can be intuitively applied to describe many kinds of visual events. Based on CL descriptions of visual events, we propose a framework for small human group event parsing in videos based on trained probabilistic context free grammar models. We also introduce a new method to describe spatial and temporal semantics for grammar induction. For the spatial semantics, both individual actions and small group behavior are taken into account for visual event representations. As to temporal semantics, the dynamic structures of multiple human objects in the scene are captured all the time, which ensures the SCFG to construct precise representations of all the targets.

As shown in Figure 1, a particular visual event is represented with five primitives. A "path" is composed of a sequence of places. A "place" is associated with the exact location and time duration a particular thing. "Things" can be human or objects, depending on different scenarios. "Action" is the the corresponding action of the thing, which can be treated as visual primitive events in most aforementioned methods. A "cause" can a special event or object causes other event or object to occur.

Unlike event-driven methods, the grammar learning is performed at two-levels of things-driven. In the first step, different things will be merged based upon their semantic distance. The mering process will be continued until it

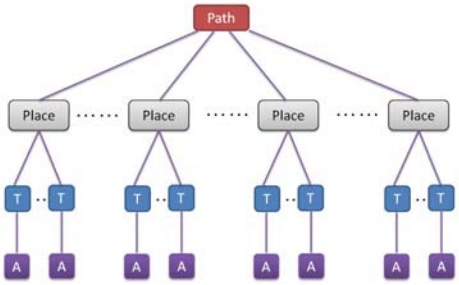

Fig. 1. Semantic structure representation

reaches the minimum semantic disorder which is defined in [11]. If multiple things have been merged as a group of things in the current place, a small group event recognition [12] will be performed and all human objects in the place will form a new thing, i.e.“group”, and its corresponding action will become small group action. If there are multiple groups, each group will be processed by the same procedure. If there are still individual persons outside the groups, they will maintain their individual descriptions.

After things merging, all the related concurrent events in the same place will be processed to form another high-level semantic representation, which avoids the parallel sub-event difficult in sequential grammar systems. The mixed descriptions of individual and group primitives will be used in the training of stochastic context free grammar rules, a minimum description length (MDL) based grammar induction method will applied to the event sequence and different rules will be generated. The induced grammar rules will be used to parse different videos. It should be noted that, in this work, if there are individual and group co-existed in the place, we treat them as non-related concurrent events and will process them separately.

The main contributions of this work can be summarized as follows:

- We propose a general visual event description framework based on cognitive linguistic primitives.
- We propose a two-level grammar induction algorithm, which can perform both individual and group event recognition, and effectively address multiple concurrent events in the scene. Unlike predefined grammar rules, all the rules are automatically induced in our method, which makes our framework adaptable to different scenarios.

2. PROPOSED FRAMEWORK

Our proposed framework is focused on the small human group action recognition. Firstly, each human object’s semantic information will be used for small human group detection. Once a small human group is discovered, their action will be recognized based on group action classification method, otherwise each individual’s action will be classified based on single human action recognition method.

Once these human action recognition results are obtained, a probabilistic context free grammar model will be initialized to automatically induce the potential rule behind group or individual action atoms. The induction will take into account both spatial and temporal correlation of each action and generate a number of events to represent different combination of actions. These induced event rules then will be used to parse different testing videos.

2.1. Cognitive Linguistics Representation

As summarized in [10], cognitive linguistics is an emerging theory of human language acquisition, and it claims that a semantic concept can be described using five conceptual primitives, i.e.“things”, “places”, “paths”, “actions” and “causes”. Within visual semantic domain, these five primitives can be interpreted as the following:

- *Paths:* Path consists of a sequence of places. Things and actions are moving along paths.
- *Places:* Places are particular locations with some time duration. Things and actions occur in places.
- *Things:* Things can be human or objects. They will conduct certain actions.
- *Actions:* Actions are dynamic behaviors of things, such as walking, running, and fighting.
- *Causes:* Causes are the reasons of different actions, which are represented as a set of functions.

More specifically, path provides an abstract description of activities, such as group meeting, group fighting, etc, over a certain space and time. Place indicates the beginning and ending frame of the event, as well as the location of each agent in the group. Things refers to group members, while action is his/her activities in the place. Given an input video with ground truth detection and tracking results, we divide it into small clips with one second, each second is a place, and the all the humans in the scene are things, their movements are the actions. Therefore we transfer the whole video to our cognitive linguistic description with in a 5-tuple representation. For example, $\{WalkTogether, < x, y, w, h, t > , < person1, walking, person2, walking >\}$ shows there are two people walking together.

2.2. Things Merging

Inspired by the social networks entropy defined in [11], we define similar criteria for things mering in the cognitive linguistic representation. Given a group G of N things, the group entropy $H(G)$ is defined as:

$$H(G) = -\sum_{i=1}^{N-1}\sum_{N}^{j=i+1} s_{ij}ln(s_{ij}) + (1-s_{ij})ln(1-s_{ij}) \quad (1)$$

where s_{ij} is the similarity measurement of two different things. In our case, each thing has a semantic description, which includes speed, direction, action, and location. s_{ij} is a measure of the semantic distance between two things. It should be noted that the things group will reach the maximum entropy value when all the things are grouped into the same cluster.

```
Input:$P$, $imax$
$H_0 \leftarrow H_P^0, i \leftarrow 0$
while $i \leq imax$ do
    i ← i+1 A,B ← random clusters from $P$
    x ←random thing from A, B, $x \in A$ move x to B,
    B ← x
    $H_i \leftarrow H_i^P$
    if $H_i > H_{i-1}$ then
        move x back to A, A ← x
    end
end
Output: $P_H$, new groups with a reduced entropy
```

Algorithm 1: Entropy Based Things Merge

The merging algorithm 1 require a initial group partition, here we used the minimum span tree to obtain the first group set based on their topological distribution. To evaluate our algorithm, we compare our semantic things merging algorithm with the method proposed in [13] in the Table 1:

2.3. Hierarchical Human Group Action Grammar Rule Induction

Grammar induction has been studied for many decades. Minimum description length (MDL) [14] has been widely accepted as the criterion for grammar induction. The minimum description length principle is define as:
1. the length, in bits, of the description of the theory
2. the length, in bits, of the data when encoded with the help of the theory.

Table 1. Small human group detection

data set	our detection rate	detection rate [13]
SU1	60.4	55.4
SU2	71.5	64.6

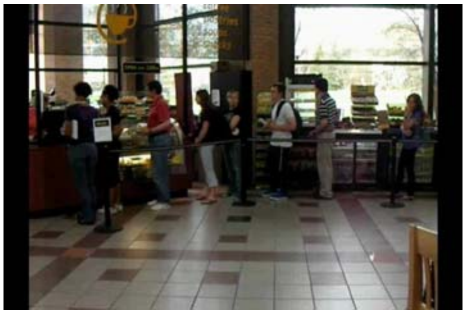

Fig. 2. A group of queuing people

As shown in [15], grammar induction process iteratively performs the merge and construct operations on the training text, until it reaches the minimum description length. We follow the description length defined in [15],

$$DL(t^L) = DL(t^L|G) + DL(G) \quad (2)$$

where t^L is the text sequence, G is the grammar, and DL is the description length.

1. *Merge* For each action pair *(A,B)*, the merge operation will create a new candidate $P \rightarrow A|B$.

2. *Construct* For each action pair *A,B*, the construct operation will produce a set of candidates, $P \rightarrow AB$.

We adopted the basic procedure for visual events grammar induction. However, due to the complexity of visual events, we extended it to a semantic merge operation in our system. The basic merge operation is based on information theory, which compressed the event symbol sequence based upon entropy from signal process perspective. This principle is effective for grammar induction from text. However, as visual event primitives have much rich information besides of symbol itself. Some work has been done in this direction [7, 8] from the spatial and temporal similarity of different events. Beside these similarity, we propose a semantic merge operation, which merge the things based on semantic representation. Take the group action "queuing" for an example, as shown in Figure 2, if there are eight people in a queue, the basic description are eight concurrent queuing action primitive, or eight stand-move events in a bottom-up description framework. As these are repetitive events, the basic merge operation will simply merge it to just one stand-move event, which will lead to a misrepresentation.

In our framework, we will apply a semantic merge operation of things as shown in the previous section. The basic idea is trying to find the most descriptive information which close to natural human language. In the "queuing" example, after semantic merge the system will output "a group of people is queuing" to describe such scene, which is more acceptable for human understanding. To do this, we firstly perform things merge at each place, if there is a human group, the group action will be classified based on [12].

The procedure of grammar induction is shown in algorithm 4 below:

```
Event symbol sequence
Grammar rules
initialization
while $\delta(descriptionlength) > 0$ do
    human group detection
    update event description
    semantic merge
    calculate description length
end
Output leaned grammar rules
```

Algorithm 2: Hierarchical Grammar Induction

Table 2. Group Primitives

symbolic definition	frequency	semantic
s1	0.2735	In Group
s2	0.0085	Approach
s3	0.1880	Walk Together
s4	0.1709	Split
s5	0.1624	Following
s6	0.0085	Chase
s7	0.0427	Ignore
s8	0.0171	Fight
s9	0.0769	Run Together
s10	0.0513	Meet

As to the video event primitives, in this work, we define three different primitive for individuals: “walk”, “run”, and “stand”. We also use ten different primitives for group actions from $BEHAVE$ data set: InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase , Fight, RunTogether, and Meet.

2.4. Video Parsing

After extracting low-level features and performing classification, each individual’s action primitive will form a string for parsing. We utilize the Earley parser [16] for parsing. The video event parsing can be iteratively processed through three steps: prediction, scanning and completing.

1. *prediction* A list of possible states will be generated based upon previous input.
2. *scanning* During scanning, the similarity between derived symbol and input string will be evaluated.
3. *completing* Based upon states selected from scanning step, the completing step will update all the positions for the pending derivations.

3. EXPERIMENTAL RESULTS

Our experiments have been conducted on two video data sets.

Table 3. Learned rules

production rule	probability	semantic
$P_{14} \rightarrow s_2 s_3$	1	group walk together
$P_{12} \rightarrow P_{11} s_4$	1	group crossing
$P_{16} \rightarrow S_9 P_{15}$	1	group fight and split
$P_{18} \rightarrow P_{13} s_5$	1	group following
$P_{26} \rightarrow P_{11} S10$	1	group meeting,talking

Table 4. Recognition result on the Collective Activity data set

group actions	detection rate	detection rate [18]
Crossing	70.4	55.4
Walking	71.5	64.6
Waiting	60.4	63.3
Talking	61.5	57.9
Queuing	90.3	83.6

3.1. BEHAVE data set

The BEHAVE data set[17] consists of four video clips, and $76,800$ frames in total. This video data set is recorded at 26 frames per second and has a resolution of 640×480. Different group atom activities include: InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase , Fight, RunTogether, and Meet. There are 174 samples of different group activities in this dataset as shown in Table 2.

After processing with algorithm 4, some of the learned probabilistic grammar rules are shown in the Table 3.

3.2. Collective Activity data set

Collective Activity dataset [18] contains 5 different collective activities : crossing, walking, waiting, talking, and queuing and 44 short video sequences. Unlike BEHAVE data set, all the videos in this data set are recorded from real-world scenarios instead of controlled environment. We use the grammar rules learned from BEHAVE data set to parse videos with different group activities. The recognition result is shown in Table 4. Comparing to the benchmark result [18], the proposed framework clearly demonstrate its advantage over the feature based methods, as some human group activities clear have state transition, which is suitable for our proposed framework.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a general representation framework for visual events based on cognitive linguistics. The general representation will then be used in an automatic grammar induction system for human group event recognition. The proposed framework successfully address the concurrent subevent problem in grammar systems. The semantic merge operation ensures that the representation is close to the human language as well as with a minimum description length. The experimental results in small human group event recognition applications demonstrate the effectiveness of our proposed framework. Although we utilized some domain-specific knowledge in this application, the general representation can be applied to many different visual recognition systems.

5. ACKNOWLEDGMENTS

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Grant FA8650-11-1-7152.

6. REFERENCES

[1] M. S. Ryoo and J. K. Aggarwal, “Recognition of composite human activities through context-free grammar based representation,” in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, Washington, DC, USA, 2006, pp. 1709–1718.

[2] D. Minnen, I. Essa, and T. Starner, “Expectation grammars: leveraging high-level expectations for activity recognition,” in *Computer Vision and Pattern Recognition, 2003 IEEE Computer Society Conference on*, jun. 2003, vol. 2, pp. 626–632.

[3] Vlad I. M. and Larry.S. D, “Multi-agent event recognition in structured scenarios,” in *Computer Vision and Pattern Recognition, The 24th IEEE Conference on*, Colorado Springs, CO, USA, jun. 2011, pp. 3289–3296.

[4] Z Si, M Pei, B. Yao, and S Zhu, “Unsupervised learning of event and-or grammar and semantics from video,” in *Computer Vision, 2011 IEEE International Conference on*, Barcelona, Spain, nov. 2011, pp. 41–48.

[5] D. Moore and I. Essa, “Recognizing multitasked activities from video using stochastic context-free grammar,” in *Eighteenth national conference on Artificial intelligence*, Edmonton, Alberta, Canada, 2002, pp. 770–776.

[6] S. Joo and R. Chellappa, “Attribute grammar-based event recognition and anomaly detection,” in *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, jun. 2006, pp. 107–114.

[7] Z. Zhang, T. Tan, and K. Huang, “An extended grammar system for learning and recognizing complex visual events,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 240–255, 2011.

[8] B. Jin, W. Hu, and H. Wang, “Human interaction recognition based on transformation of spatial semantics,” *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 139–142, mar. 2012.

[9] R. Jackendoff, *Semantic Structures*, Mit Press, 1993.

[10] J. Mitola and H. Man, “Semantics in cognitive radio,” in *Semantic Computing,(ICSC). IEEE International Conference on*, sept. 2009, pp. 261–266.

[11] J.D. Cruz, C. Bothorel, and F. Poulet, “Entropy based community detection in augmented social networks,” in *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, oct. 2011, pp. 163–168.

[12] Y. Yin, G. Yang, J. Xu, and H. Man, “Small human group activity recognition,” in *Image Processing, the 2012 IEEE International Conference on*, Orlendo, FL, USA, oct. 2012, pp. 2709–2712.

[13] W. Ge, Collins R. T., and Ruback R. B., “Vision-based analysis of small groups in pedestrian crowds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.

[14] Jorma Rissanen, “A universal prior for integers and estimation by minimum description length,” *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, 1983.

[15] Grnwald Peter, “A minimum description length approach to grammar inference,” in *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*. 1995, vol. 1040 of *Lecture Notes in Computer Science*, pp. 203–216, Springer.

[16] A. Stolcke, “An efficient probabilistic context-free parsing algorithm that computes prefix probabilities,” *Computational Linguistics*, vol. 21, no. 2, pp. 165–201, jun. 1995.

[17] S. Blunsden and R. B. Fisher, “The behave video dataset: ground truthed video for multi-person,” *Annals of the BMVA*, vol. 4, pp. 1–12, 2010.

[18] W. Choi, K. Shahid, and S. Savarese, “What are they doing? : Collective activity classification using spatio-temporal relationship among people,” in *IEEE 12th International Conference on Computer Vision Workshops*, Kyoto, Japan, oct. 2009, pp. 1282–1289.