ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/308980342

Event-driven Movie Annotation using MPII Movie Dataset

Conference Paper · October 2016

CITATION
1

READS
139

1 author:

Tuan Do
Brandeis University
5 PUBLICATIONS **8** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Teaching virtual agent to perform actions View project

All content following this page was uploaded by Tuan Do on 11 October 2016.

The user has requested enhancement of the downloaded file.

Event-driven Movie Annotation using MPII Movie Dataset

Tuan Do

Brandeis University

Abstract. This paper presents an ongoing work that aims to create a dense layer of annotation on top of the MPII Movie Description Dataset (MPII), focusing on events and event participants. In this work, we select a small set of movies from MPII and using an annotation tool (Event Capture Annotation Tool) to annotate events over some temporal span and location, posture or movement of their participants. We would also analyze the task's dimensions by giving some statistics on the resulted annotation data (at the time of submission). We believe that the work would be a valuable resource for any future work regarding event understanding and interpreting from movies and videos in general.

Keywords: movie annotation, event annotation

1 Introduction

Video understanding is a challenging problem. It requires state-of-the-art computer vision algorithms and techniques to segment and recognize objects and their attributes out of the image background, tracking them over time for their movements, resolving their interactions and finally making sense of these information. To match these understandings with some textual descriptions would require another set of algorithms, including but not limited to selection of salient events and event participants and generation to natural language. Without a sufficiently annotated dataset, these would build up to insurmountable difficulties. For example, research on image understanding has made great progress over recent years, leveraging on large and semantically rich datasets of images such as Flickr30k [1], Microsoft CoCo [2] and Visual Genome [3]. However, because of the temporal dynamic of videos, there are fewer publicly available video and movie resources with rich semantic annotations. To the best of our knowledge, MPII Movie Description [4] is the first extensive dataset that has parallel movie snippets and descriptions. These descriptions are transcribed from audio description for visually impaired people. Therefore, they are highly event-centric, describing the most salient event(s) in each movie snippet. Let's take a look at some snippet descriptions, compared to some rough translations of what happens in the corresponding snippets, taken from the movie "Forrest Gump":

1. *Audio description* He opens a box of chocolates and holds it out for the nurse. *Visual description* One man and one woman are sitting on a bench

in front of an open public lawn, the woman is reading a magazine while the man is looking at his box of chocolates. He opens the box and holds it out for the woman. The woman then looks at the man.

2. *Audio description* The nurse shakes her head, a bit apprehensive about this strange man next to her. *Visual description* The woman looks at the box, shakes her head, then continues reading her magazine. The man keeps holding the box until the end of the snippet.

There are some interesting phenomena to note here. One thing is that the most salient event(s) in each snippet normally constitute only some parts of the snippet. Other events that take considerable amount of time might be irrelevant to the progress of the narrative. However, image processing algorithms might not be able to distinguish between salient events and non-salient events. Context understanding across some contiguous snippets is also required. For example, visual and textual co-references resolution need to be established between the woman seen in both snippets and mentions of "the nurse" in the audio description. Moreover, narrative understanding also requires understanding of causality relationship between events as well as their novelty; without this information, the most salient event of the second snippet is probably that the man offered the chocolate to the nurse.

We believe that multiple layers of semantic annotation could be added to this dataset to provide the information needed to facilitate understanding. Firstly, annotators mark spanning of described events. Secondly, annotators mark objects participating in these events, linking them to their corresponding textual expressions. They give them names in a descriptive manner, such as "old man", "boy", rather than give them a semantic type drawn from a fixed ontology. Lastly, events in the form of unary or binary relationships between objects are added to either describe the posture or movement of objects (such as STAND, WALK, RUN) or describe certain interactions among them (such as SUPPORT, LOOK_AT). Later on, we will give more details on why we choose these specific event types to annotate.

2 Related work

Image annotation Extensive images datasets with various level of semantic annotation have been released in recent years, such as Flickr 30k [1], MS-COCO [2] and Visual Genome [3]. While the first two datasets give much focus on annotation of objects, by using either bounding box or segmentation, the last dataset leverages object relationships and attributes to first-class citizenship. We adopt this view in our annotation work, giving considerable effort to annotate object relationships and attributes. However, in our framework, we do not annotate the typical object attributes, such as shape, color, size, but we focus on event or action predicates such as $STAND$, RUN, $LOOK\_AT$.

Textual event annotation Textual event annotation has been covered extensively in a number of studies, such as TimeML project [5] and NarrativeML [6].

Events in the textual context are quite different from events in movie annotation. For instance, events in TimeML have tense and aspect and are anchored by relationship with time mentions in narratives. We do not annotate them in our annotation framework. The reason is partly because of the fictional and possibly nonlinear nature of movies, leading to difficulty in annotating event aspects and anchoring events with any timescale.

Movie annotation There are not many available video or movie resources that have similar level of semantic annotation as targeted in our work. For its object detection from video task, the Large Scale Visual Recognition Challenge 2016 (LSVRC) released a dataset with categorical objects annotation [1]. Mani [7] suggested an annotation framework for cartoon employing topological relations but does not release any annotation dataset.

3 Framework

3.1 Event

Free-form event Free-form event is represented by a textual description, with event participant mentioned and linked to annotated objects. For example, taken the description of the first snippet in Forrest Gump, we separate it into two textual events: *He opens a box of chocolates* and *He holds it out for the nurse.* Each event will be marked with a duration in the snippet $[f_1, f_2]$ and mentioned of "he", "box of chocolate", and "the nurse" will be made references to annotated objects.

An important point to mention about visual representation of events is that they might or might not encompass the whole progress of described events, or they might only describe the result states only. Let's take a scene showing a team of constructors building a new house as an example. Technically speaking, a "build" event is of accomplishment event type ([8, chapter 6, p. 182]), involving a preparatory phase, a culminating point and a result state. This visual scene, however, only involves the preparatory phase. Another scene shows the house completed with a description like "The house is built". In this case, the description is not about the "build" event but about its result state only. We leave out any description that only describes a result state and mark spanning of events of achievement and accomplishment types to the culminating point only.

Structured-form event Structured-form event is represented by a proposition that holds for an interval of time, such as $RUN(object, t) \wedge t = [f_1, f_2]$. The way annotators add propositions is going to a specific frame f in the video snippet, select an object, and select one of the predefined predicates. By default, the time interval has infinite endpoint $t = [f, \infty]$. When the proposition no longer holds, annotator needs to set its truth value to $False$ or add a conflicting proposition

[1] http://vision.cs.unc.edu/ilsvrc2015/ui/vid

at a later frame (constraints on propositions are discussed shortly) to close the interval. We predefine a set of unary predicates, such as $STAND$ and a set of binary predicates such as $LOOK\_AT, POINT\_AT$. It is worth noting that the same annotation routine could also be used for intrinsic attributive relationships between objects, such as $PART\_OF$, or spatial relationship.

The full list of predicates in structured-form events are:

- Unary predicates: $LIE, SIT, STAND, CROUCH, KNEEL, WALK, RUN$. These primitive postures or movements per se are already informative about what happens in a snippet. More importantly, changes between these postures or movements signify other actions or events. A change from walking to running of a person might signify a $CHASE$ or $CATCH$ action while a change from standing to kneeling might signify a $PRAY$ action.
- Binary predicates: $LOOK\_AT$, $POINT\_AT$, $SUPPORT$, $PART\_OF$, $IDENTITY$ Inclusion of $LOOK\_AT$ and deixis $POINT\_AT$ are motivated by the fact that these are important actions that signify conversational interaction. Both of them could also be used to introduce new events and event participants in narratives. $SUPPORT$ is used to indicate an object resting on another object and moving together with it. It could be used for numerous cases, such as $SUPPORT(car, person)$ when a person is riding a car, $SUPPORT(mom, child)$ when a mother is carrying her child and $SUPPORT(person, shoes)$ when a person is wearing a pair of shoes. $PART\_OF$ relates two part-whole objects. $IDENTITY$ is to link inter-snippet object annotations that refer to the same visual object.

We specify some constraints in the form of mutual exclusion of propositions on a same argument. For example, to specify that one person could not run and sit at the same time, we add the following constraint into annotation tool (notations from Temporal Interval Logic [9]):

$$RUN(o, t_1) \wedge SIT(o, t_2) \implies t_1 \bowtie t_2 \tag{1}$$

This constraint automatically closes the infinite interval of the first proposition in the following case:

$$RUN(o, t_1) \wedge t_1 = [f1, \infty] \wedge SIT(o, t2) \wedge t2 = [f2, \infty] \wedge f1 < f2 \implies t_1 = [f1, f2] \tag{2}$$

Based on annotators' inputs, we found that adding these constraints ease the annotation process by automatically checking consistency of added events.

3.2 Object

Objects are annotated by drawing bounding box at some frames in the snippets, similar to annotation in LSVRC. However, we do not aim to annotate exact locations of objects at all frames. Instead, by assuming that object bounding boxes move smoothly over some period of time, annotators could pick some

critical frames to annotate objects' exact locations. For frames in between, their bounding boxes are linearly interpolated.

It is worth emphasizing that we do not have an ontology for object semantic types. The main reason is that the natures of objects in the movies are unknown beforehand. Annotators do not need to study the complexity of a full-fledged ontology, such as Wordnet. Instead, annotators give names to objects so that they can easily remember them.

4 Annotation

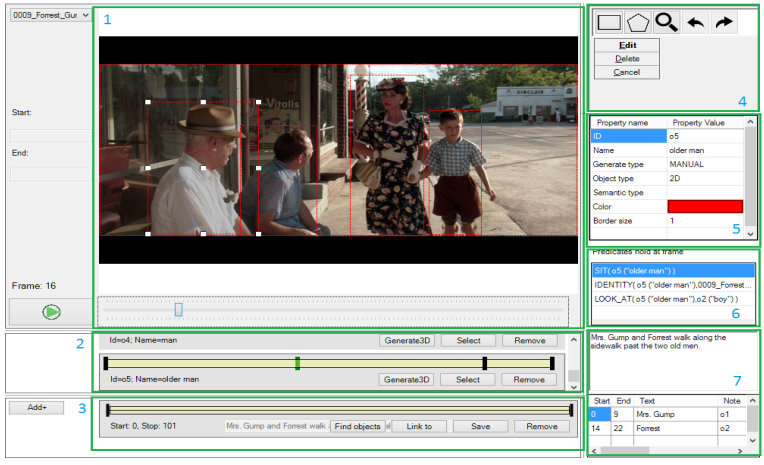

Fig. 1. A sample of annotation GUI as seen when a snippet of Forrest Gump is being annotated. (2) shows temporal span of objects, black and green markings are frames that annotators add their locations and relations. (3) shows temporal span of free-form events. (5) shows a selected object's properties. (6) shows relations of the selected object's at a specific frame. (7) shows textual references from description to annotated objects

We selected the following 10 movies from MPII dataset:

We gave the task to 2 annotators who are undergraduate students and have intermediate English skills. Annotation are carried out in two phases:

- Phase I: We first gave annotators tutorial videos, then we selected 10 snippets in each movie and gave to annotators to independently work on them (for a total of 100 snippets). We also created a gold annotated dataset on the same number of snippets. By getting feedback from annotators, we made modifications to annotation tool and completed the annotation guidelines (completed)
- Phase II: Each annotator annotates a whole movie. (ongoing)
- Phase III: Analyze annotator inter-agreements using double-annotation movies. Dataset (annotation files without movie snippets) will be released for public access (planned)

Table 1. List of movies. The movies are selected based on the criteria of being diverse in genres, of high quality and having generally bright tone. Names in bold are movies that we aim to provide double annotation

ID	Movie name	# snippets	Genres
0009	**Forrest Gump**	910	Drama, Romance
0016	O Brother Where Art Thou	503	Adventure, Comedy, Crime
0017	Pianist	815	Biography, Drama, War
0046	Chasing Amy	357	Comedy, Drama, Romance
0051	**Men in black**	439	Adventure, Comedy, Family
1028	No Reservations	608	Comedy, Drama, Romance
1033	Sherlock Holmes A Game of Shadows	838	Action, Adventure, Crime, Mystery, Thriller
1043	Vantage Point	634	Crime, Drama, Mystery
1051	Harry Potter and the goblet of fire	898	Adventure, Family, Fantasy, Mystery
1060	Yes man	591	Comedy, Romance
Total		6593	

4.1 Annotation tool

We adapted a tool named ECAT ([10]), which is originally used for capturing and annotating events with three dimensional data. ECAT supports marking events, objects and binary relations between objects. We made the following modifications to the tool for our work:

- Add a view for annotator to track propositions held at a certain time frame.
- Add inter-session relations of objects. That allows us to add relation $IDENTITY(o1$ for two annotated objects in different sessions that refer to the same visual object.
- Modify predicate settings to include our constraint mechanism.

4.2 Other Important Points of Guidelines

- Bounding boxes of a person should fit his head, torso and legs. They should be added at the start and end frame of the snippet if the object persists; otherwise, annotators need to indicate when the object disappears out of the scene. If the resulted interpolation does not fit object locations properly, add more bounding boxes at the points which interpolated results differ the most.
- A quick way to annotate objects if different snippets describe continuous progress of the same scene is to copy the objects between the snippets. Each cloned one has an $IDENTITY$ relation with its corresponding original one.
- If there is a group of people or objects and there is no reason to separate between them (like audiences in a football stadium), annotate them entirely as one object.
- Only after objects are annotated, structured-form events are added.

4.3 Dataset

At the time of submission, we have gathered annotations for around 900 snippets (of two movies Forrest Gump and Men in black), though we have not collected enough for double-annotation analysis. Followings are some statistics of these annotations, which give an estimation of the dataset dimension at the time of completion:

Table 2. Some statistics of annotated data

	Value
Number of completed snippets	909
Number of objects	3365
Average number of objects per snippet	3.7
Number of object location markers	12358
Number of location markers per object (higher mean finer granularity)	3.67
Number of structured-formed events	3076

Table 3. Breakdown of structured-form events

Pred	Value	Pred	Value	Pred	Value	Pred	Value
LIE	103	SIT	420	STAND	790	CROUCH	85
KNEEL	16	WALK	407	RUN	212	SUPPORT	226
LOOK_AT	784	POINT_AT	33				

5 Future considerations

The first consideration is to include object types into our dataset. We might take advantage of some available word vector similarity tools to match objects' names from our annotated data to a proper semantic type. The result would be double-checked by an expert.

Another consideration is to match predicates of events to semantic frames such as Framenet [11]. That would allow us to specify semantic roles of event participants in a systematic way, especially for free-form events.

Event attributes could also be added in future extension of our framework. One of the reasons is to distinguish between different types of *eventualities* as discussed previously in 3.1. Currently we do not have any method to specify this information.

6 Conclusion

Our annotation work is the first attempt to bridge the gap between visual scenes in movies and their audio descriptions. Our annotation framework includes, but not limited to, marking spanning of events mentioned in descriptions, marking event participants, marking their postures and interactions, and marking participant co-references between snippets.

We hope that our annotated dataset at the time of completion would be a helpful resource to facilitate research in the language and vision communities. Moreover, it could serve as a starting point for future semantic annotation studies leveraging on both visual and textual data.

References

1. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2** (2014) 67–78
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, Springer (2014) 740–755
3. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332 (2016)
4. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. (2015)
5. Pustejovsky, J., Castano, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G., Radev, D.R.: Timeml: Robust specification of event and temporal expressions in text. New directions in question answering **3** (2003) 28–34
6. Mani, I.: Computational modeling of narrative. Synthesis Lectures on Human Language Technologies **5**(3) (2012) 1–142
7. Mani, I.: Animation motion in narrativeml
8. Cann, R., Kempson, R.M., Gregoromichelaki, E.: Semantics: An introduction to meaning in language. Volume 198. Cambridge University Press New York (2009)
9. Allen, J.F., Ferguson, G.: Actions and events in interval temporal logic. Journal of logic and computation **4**(5) (1994) 531–579
10. Do, T.: Ecat: Event capture annotation tool. Workshop on Interoperable Semantic Annotation (2016)
11. Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R., Scheffczyk, J.: Framenet ii: Extended theory and practice (2006)

View publication stats