Annotating News Video with Locations

Jun Yang and Alexander G. Hauptmann

School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{juny, alex}@cs.cmu.edu

Abstract. The location of video scenes is an important semantic descriptor especially for broadcast news video. In this paper, we propose a learning-based approach to annotate shots of news video with locations extracted from video transcript, based on features from multiple video modalities including syntactic structure of transcript sentences, speaker identity, temporal video structure, and so on. Machine learning algorithms are adopted to combine multi-modal features to solve two sub-problems: (1) whether the location of a video shot is mentioned in the transcript, and if so, (2) among many locations in the transcript, which are correct one(s) for this shot. Experiments on TRECVID dataset demonstrate that our approach achieves approximately 85% accuracy in correctly labeling the location of any shot in news video.

1 Introduction

Annotating the geographical location of video scenes is a critical step towards semantic video analysis and retrieval. However, there has been very limited research on this problem [1,3,6]. The goal of this paper is to automatically annotate the location of every shot in broadcast news video. Achieving this goal will leverage high-level retrieval tasks on news video, such as *"Find the scenes showing the flood in California caused by El Nino"*, or *"List the countries that President Bush visited last year and find the scenes of each visit"*.

There have been several efforts on labeling video with locations. One method is to use image characteristics to match the current shot against a set of existing shots with known locations, which has been used by Aoki et al. [1] and Sivic et al. [8]. However, it has limited applicability in news video because the footage contains a huge number of locations with diverse scenes for each one, making the collection of example shots for every location impossible. A separate track of research has used GPS information to determine location [6], which is not available for news video. Christel et al. [3] have successfully used locations extracted from the transcript of news video to create an map-based interface for browsing, but they did not correlate the locations with specific shots. To our knowledge, there is no working approach for annotating the locations of news video shots.

The *general* problem of annotating the locations of video of arbitrary genres is extremely difficult. The *specific* problem we are focusing on, namely annotating locations of broadcast news video, is tractable because news video comes with transcript from closed-captions or speech recognition, which contains most

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 153–162, 2006.

*... fray between the **United States** and **Iraq** ... **U.N.** secretary general Kofi Annan will go to **Baghdad** ... tanks were training in the sands of **Kuwait** ... meeting five permanent members of **U.N.** security council, the **U.S.**, **Russia**, **China**, **France**, and **Britain** ... flexibility by **Iraq** in allowing weapons inspectors ...*

Fig. 1. A sequence of video shots from a news story and the locations in transcript

of the locations shown in the footage. Nevertheless, this specific problem is still challenging for several reasons. First, there are typically more than one location mentioned in the vicinity of each shot, and the true location of the shot is not necessarily the closest one. Second, determining the location from the visual content of a shot is virtually impossible, because one location can have numerous visually different scenes. Last but not the least, some shots do not have a legitimate location, such as the shots showing stock market data, and some have locations that are not worthwhile to be mentioned, such as anchor shots. It is nontrivial to tell if the location of a shot is among those in the transcript.

These difficulties are illustrated in Figure 1, which shows a news story on the Iraqi crisis in 1998, where the locations of the footage switch between Kuwait, United Nations, and Iraq. One difficulty is that the order in which the locations appear in the transcript is different from the order of the shots showing these locations. Moreover, one has to get rid of extra locations such as Russia, China, and France, which are mentioned in the transcript but never shown in the footage. Finally, one needs to tell that the location of the anchor shot is not among those mentioned in the transcript.

As parallel streams of information, *correlations* exist between the mentions of locations in the transcript and the changes of the video scenes to ensure the footage being comprehensible. In this paper, we capture the location-shot associations by exploring clues from different modalities of the news video, including the syntactic analysis of the transcript, temporal video structure, speaker identification, and so on. Machine learning methods are adopted to combine these multi-modal features to solve two sub-problems: (1) is the location of a given shot mentioned in the transcript? and if so, (2) among the many locations in the transcript, which are the correct location(s) of the shot? Experiments on TRECVID dataset demonstrate that our approach achieves 85% accuracy in correctly labeling the location of any shot in news video.

2 An Overview of the Approach

News video footage consists of a series of stories, where each story is a semantically coherent video sequence on a specific news event. A story can be further partitioned into shots, and each shot contains the scene at a specific location.

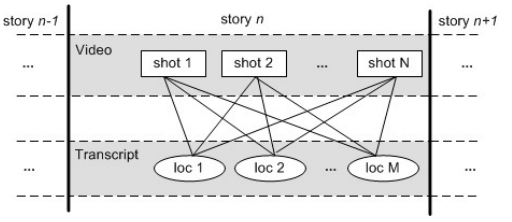

Fig. 2. The formulation of location annotation in news video

Automatic segmentation of stories and shots can be done with high accuracy. Moreover, we can obtain the transcript of news video from its closed-captions (CC) and/or using automatic speech recognition (ASR). All the mentions of locations can be extracted from the transcript (Section 3). ASR text is temporally aligned with the video during its generation process, while CC text can be aligned to video by matching it with ASR text. Thus, the time-stamp of every mention of location in the transcript is known.

As each story is an independent unit, the location of a shot (if mentioned) needs to be searched only among the locations appearing in the transcript of the same story, known as the *candidate locations* of the shot. Figure 2 suggests that location annotation is about finding the correct associations between shots and locations within the boundary of each story. Specifically, we can predict the location of $Shot_i$ by evaluating its probability of being associated with each of its candidate locations $\{Loc_{ij}\}$, denoted as $P(Match|Shot_i, Loc_{ij})$. Each shot-location association is described by a set of multi-modal features that help distinguish the correct/incorrect associations, as will be elaborated in Section 4. Once the probabilities are computed, we can annotate the shot with the location(s) with high probability. Note that one shot can have more than one locations, e.g., California and San Francisco are both valid locations for a shot showing San Francisco. On the other hand, the locations of some shots never appear in the transcript for various reasons, an issue to be further discussed in Section 5.

This formulation leads to a supervised binary classification problem of distinguishing correct and incorrect shot-location associations. Using any existing learning model, we can learn a classifier from example shots that have manually labeled locations, and then use the classifier to predict the probability of each unlabeled shot being associated with each of its candidate locations. We explore two learning approaches in our experiment, namely logistic regression and support vector machine (SVM).

3 Extracting Candidate Locations

The candidate locations are automatically extracted from the video transcript using the BBN named-entity detector [2]. From its output, we take all the terms/phrases recognized as "location" as our candidate locations. Additional

locations are mapped from "organization" terms/phrases with self-contained locations, such as *"Capitol Hill"*, using a manually created mapping list. Note that location terms are sometimes superimposed on the video frames, which can be recognized by video optical character recognition (VOCR) techniques [7]. However, the VOCR output tends to be errorful on low-resolution news video, and they offer few *distinct* locations since most of them overlap with those from transcript. Thus, currently we do not include these locations as the candidates, and leave it for future research to utilize such errorful locations.

Two problems need to be addressed to transform the extracted raw locations into those used for annotation: *location synonymity* and *location polysemy*. The synonymity problem arises when there are multiple representations of the same physical location, which can be caused by abbreviations, such as *"NY"* and *"New York"*, specificity, such as *"Long Island"* and *"Long Island, New York"*, canonical names and variants, such as *"Holland"* and *"Netherland"*, etc. By looking up each location term in a a geographical dictionary, or a gazetteer[1], we merge synonymous locations to create a set of distinct candidate locations. The gazetteer has various representations of a location and the hierarchical relationships between locations, which, for example, tells the fact that *"Long Island"* is inside *"New York"*. An item of the gazetteer looks like "***Paris*** *– French; Built up area; ...; France; Europe;*", where it shows the language, coordinate, category, and country and continent of each location.

In contrast, the polysemy problem refers to the case where two or more different physical locations share the same representation. For example, *"London"* can be a city in United Kingdom or a city in Ontario, Canada, and if appearing by itself, it is impossible to tell which city is referred to. We disambiguate such polysemantic location terms by considering the *context* information. For example, if a location term has two possible references, and we find other locations in the same story that either subsumes or is subsumed (based on the gazetteer) by one of the referred locations, we decide that this is the location actually referred to. If no such context clues are found, however, we simply pick the default reference of this location term suggested by the gazetteer.

4 Multi-modal Features for Location Annotation

Features from multiple video modalities are used for classification of correct and incorrect locations. In this section, we discuss the insight behind the use of each modality, and leave the details of all the features to Table 1.

4.1 Temporal Relationships

There is an apparent temporal correspondence between the progress of video shots at different locations and the mentions of location terms in the transcript. For example, generally the location mentioned closest to a shot is mostly likely its true location. We explore such temporal relationships from several aspects:

[1] We manually built the gazetteer from the information available at GEOnet Names Server (earth-info.nga.mil/gns/html) and U.S. Geological Survey (www.usgs.gov).

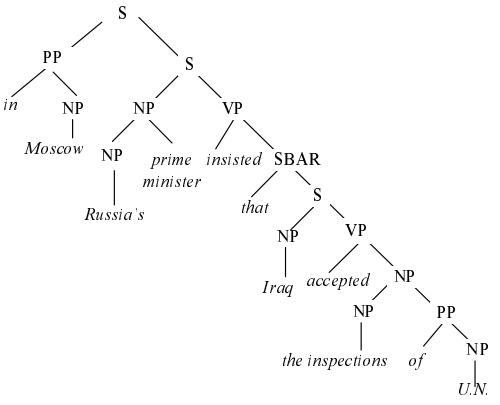

Fig. 3. Parse tree of the example sentence

Overlaid: *IRAQ*
VOCR output: *LRAQ*

Edit distances:
France: 0.67
Russia: 1.0
U.S.: 1.0
Iraq: 0.25

Fig. 4. Overlaid location

- ***order:*** whether a location is mentioned before, within, or after a shot.
- ***distance:*** the distance (in seconds) between a shot and the nearest mention of a location.
- ***closeness rank:*** how close a location term is to a shot, compared with the other locations in the same story.

4.2 Syntactic Features

The syntactic roles of a location term in the sentences of the transcript implies whether it is the actual location of the footage. For example, from sentence "*In **Moscow**, **Russia**'s prime minister insisted that **Iraq** accepted the inspections of **U.N.***", one can easily tell that Moscow is more likely the true location of the video than Iraq or U.N., since it is inside a prepositional phrase *"in Moscow"* which indicates the location of the event. The syntactic roles of a location in a sentence can be obtained from its parse tree. We use Link Grammar Parser [9] to parse sentences into parse trees. Figure 3 shows the parse tree of the above sentence, where it is decomposed into a set of nested *constituents* of several types, such as noun phrase (NP), verb phrase (VP), prepositional phrase (PP), sentence (S), sub-sentence or clause (SBAR). By analyzing the parse tree we can classify the syntactic role of a location term as one of the following:

- ***prepositional phrase*:** Video locations are often expressed via PPs, such as "*in Moscow*", so we identify all the location terms occurring in PPs. We also examine the specific preposition used in order to distinguish PPs that do not indicate locations, such as "*of U.N.*".
- ***subject/object*:** Location terms as the subject or object of a sentence are unlikely references to the actual location, such as "*Iraq*" in above sentence.
- ***modifier*:** Like Russia in *"Russia's prime minister"*, a location modifying other nouns is usually not the location of the video scene.

Table 1. The feature set describing the association between a shot S and a location L

Modality	Feature	Description
Syntactic Feature	*in-loc-pp*	L is inside a PP that indicates location
	in-other-pp	L is inside a PP that does not indicate location
	is-subj-obj	L is used as the subject/object of a sentence
	is-modifier	L is used to modify another noun or noun phrase
Temporal Relationship	*shot-loc-dist*	the temporal distance between S and L
	loc-rank	the rank of L in terms of its closeness to S
	shot-loc-order	L is mentioned before, within, or after S
Location Properties	*continent*	L is a continent
	country	L is a country
	province	L is a province or state
	city	L is a city, town, or region
	organization	L is an organization
Overlaid Text	*vocr-similarity*	the similarity between L and VOCR output of S
Speaker Identity	*anchor/reporter/ narrator/subject*	L is uttered by the anchor, reporter, narrator, or new subjects of the story

4.3 Screen-Overlaid Location (VOCR)

Location terms are occasionally overlaid on video frames to indicate the true location of the current shot. While we choose not to rely on the errorful locations recognized by VOCR [7] (Section 2), they are nevertheless useful due to their similarity to the true location terms. In Fig.4, for example, *Iraq* is recognized as *Lraq*, differing by only one character. Therefore, the string similarity between each candidate location of the shot and the VOCR output indicates which candidate matches the screen-overlaid location, and thus the true location of the shot. The similarity is measured by *edit distance*, defined as the number of insertions, deletions, or substitutions needed to convert one string into another, which is then normalized by the length of the source string. Figure 4 lists the normalized edit distances of some candidate locations to the VOCR output, where the true location *Iraq* has the shortest distance.

4.4 Speaker Identity

The identity of the person who utters a location term is also related to whether this location is shown by the video. The speaker identities of a news story include anchor, reporter, narrator, and news-subjects (i.e., people in news events). Our observation reveals that the true locations are more likely from the speech of the anchor, narrator, and reporter, who are observers of the news, rather than from the news-subjects as the insiders of the story. Speaker identification is a byproduct of the LIMSI speech recognition system [4], which groups the speech segments that are likely to be of the same speaker, with an ID assigned on each group. Although these IDs do not directly indicate the actual identity of each speaker, we can derive that from the distributions of IDs and other clues using

Fig. 5. Various types of shots without specified locations in transcript

the method described in [10]. Once the speaker identity is known, one can tell the identity of the speaker uttering each location by matching their timestamps.

4.5 Location Type

Locations of certain types are simply more (or less) likely to be the real location of a story. For example, when *"White House"* is mentioned, it is dubious whether there are footage showing the actual place, because this phrase is often used to refer to an organization, such as in *"**White House** says today that Iraq must allow the weapon inspectors."*. To capture such information, we classify locations into several types by their specificity and other properties. The type information of a location can be easily read from the gazetteer (Section 3), and is turned into a set of features as shown in Table 1.

5 Distinguishing Shots Without Specified Location

Some shots do not have a legitimate location, such as artificial shots showing maps and stock market data; some have locations but their locations do not appear in the transcript. While it makes no sense to annotate the locations of the shots in the first case, it is extremely difficult to annotate the shots in the second case since their locations can *only* be guessed from the visual content, which is beyond the start-of-the-art of pattern recognition and the focus of this paper. In our approach, we identify the shots *without* specified locations in transcript (i.e., shots in either of the two cases) and dismiss them as "unspecified", leaving the prediction of their specific locations to future work. A close examination reveals that such shots belong to the following types (1) *commercial shots*, (2) *artificial shots*, such as shots showing maps, stock market data, animations, sketches, (3) *studio-setting shots*, including anchor shots and shots showing interviews, (4) *symbolic-scene shots*, which show symbolic scenes whose locations are self-contained, and (5) *general-scene shots*, which show scenes of general types where the specific location is of no interest, such as "people at beach". Figure 5 shows examples of each type of shots.

Given the variety of video shots without specified locations, there is no simple heuristic available to identify all such shots, especially the last two types. Similarly, we formulate it as a supervised binary classification task as to distin-

guishing shots with specified locations from those without, and apply learning methods such as logistic regression and SVM to it. The features (of each shot) for this task are derived from different modalities of news video. Due to the limited space, we briefly discuss the key features below.

- ***Shot category:*** Among the aforementioned types, anchor, commercial, and weather-forecast shots can be readily identified by existing concept detectors [5] on news video, whose outputs are incorporated into the feature set.
- ***Story topic:*** Stories on business, entertainment, health, and technology are more likely to contain scenes without specified locations. Thus, we built a text classifier that predicts based on the transcript the category of each story as *politics*, *business*, *health*, *technology*, *sports*, and *entertainment*, and the predictions are incorporated as features. The classifier is trained using SVM based on news video transcript with manually assigned topic labels.
- ***Motion:*** Most artificial and studio-setting shots are close to static. Thus, we use some motion features, such as the average pixel difference between consecutive frames, to help identify such shots.

6 Performance Evaluation

Our experiment is conducted on 10-hour footage of ABC World News Tonight[2] from TRECVID 2004 collection, which consists of 6219 shots. We use a named-entity detector [2] to extract all the location terms from the closed-captions of the footage. It should be noted that our approach can also work with ASR text if closed-captions are unavailable. From the detected locations, we remove the continent names and "United States" since these general locations hardly provide any useful information. The candidate locations of each shot are the locations appearing in the same story as the shot, where the true story boundaries are provided by TRECVID. In average, each shot has 4.02 candidate locations.

To collect the truth, a human annotator gave binary judgment on whether each candidate location is correct or incorrect for a given shot. If a shot has multiple true locations with varying specificity (e.g., *"San Francisco"* and *"California"*), no ranking is enforced and they are considered equally good. If the annotator decided that a shot does not have a legitimate location, or none of the candidate locations is correct, he annotated it as *"unspecified"*. It turned out that 1768 of the 6219 shots are annotated with at least one location, with the remaining labeled as *"unspecified"*. In average, each shot has 1.41 *correct* locations out of 4.02 candidates, making the accuracy of a random annotator about 35%.

For comparison purpose, we implement three heuristic baseline approaches as benchmarks: **WindowLoc** annotates each shot with all the locations found within a temporal window (on the transcript) of 20 seconds centered around that shot, **NearestLoc** labels each shot with the temporally closest location in the corresponding story, and **MaxFreqLoc** annotates each shot with the location

[2] Due to time constraint, we are unable to experiment with other types of news video like CNN, but our approach is generally applicable.

Table 2. Performance on location annotation in two settings

Setting		Shots with specified location		All shots	
Metric		ClassAcc	LabelAcc	ClassAcc	LabelAcc
Baseline	**WindowLoc**	0.653	0.480	0.761	0.690
	MaxFreqLoc	0.712	0.576	0.626	0.518
	NearestLoc	0.712	0.641	0.626	0.513
Learning Model	**LogReg**	0.774	0.779	0.853	0.793
	SVM	**0.869**	**0.864**	0.**884**	**0.851**

that appears most frequently in the story. All the three methods annotate a shot as *"unspecified"* if no locations are found in the window or in the story.

The experiment is conducted in two settings. The first one focuses on *only* the 1768 shots with specified locations. The classifier described in Section 2 is applied to predict the probability of every shot being associated with each of its candidate locations, which can be transformed into the (correct/incorrect) labels on these locations. Two performance metrics are computed from the results of 10-fold cross-validation: Classification accuracy (**ClassAcc**) is the ratio of correctly classified candidate locations, while labeling accuracy (**LabelAcc**) is the ratio that the top-ranked candidate location of each shot (i.e., the one with the highest probability) is the correct location. This second metric is practically more meaningful since it represents the chance that users see a shot correctly labeled with at least one location. The left side of Table 2 shows the performance of five methods, including three baselines and the proposed learning methods using **LogReg** (logistic regression) and **SVM**. One can see that the proposed methods significantly outperform the baselines. SVM is the best performer, which achieve 87% accuracy on classifying locations and 86% on labeling shots. The superiority of SVM can be contributed to its RBF kernel which explores the correlations of different features. All the baselines generate results that are better than random, especially the MaxFreqLoc and NearestLoc, implying that heuristics like temporal distance and frequency are useful.

In the second setting, we use all the 6219 shots in order to evaluate our approach for identifying shots without specified locations. For each shot, we first determine whether its location is mentioned in the transcript, using a classifier described in Section 5. If the answer is negative, the shot is labeled as *"unspecified"*, otherwise we predict the location for the shot as in the previous experiment. The result showed that this pre-filtering process classifies 4072 shots as *"unspecified"*, among which only 244 are false-alarms, and it fails to identify 492 shots with unspecified locations. This suggests that our approach can distinguish shots without specified locations with high accuracy (89.7%). Treating *"unspecified"* as a special location, we show the overall accuracy of location annotation on the 6219 shots in the right side of Table 2. The proposed methods achieve 79% (LogReg) and 85% (SVM) accuracy on labeling the locations of shots. This result is very encouraging since this setting is close to the reality where a user has no idea on whether a shot's location is in the transcript or not.

7 Conclusion

This paper has presented a learning-based approach to annotate news video shots with locations based on multi-modal video features. Specifically, we have discussed and solved two problems, namely determining (1) whether the location of a given shot is mentioned in the transcript, and (2) among the locations in the transcript, which are the correct location(s) of the shot. The experiments on TRECVID dataset have shown that our approach can correctly annotate about 85% of the shots with their locations. In future, we plan to evaluate our approach on video data with ASR text to study how imperfect transcript will affect its performance, and include the locations appearing in VOCR text as possible labels of shots. Another challenging future work is to investigate the difficult task of annotating shots whose true locations are not mentioned in the transcript.

Acknowledgement

This work was supported in part by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037, and by the National Science Foundation under Grant No. IIS-0535056.

References

1. H. Aoki, B. Schiele, and A. Pentland. Recognizing personal location from video. In *Workshop on Perceptual User Interfaces*, pages 79–82, 1998.
2. D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *Proc. 5th Conf. on Applied Natural Language Processing*, pages 194–201, 1997.
3. M. Christel, A. Olligschlaeger, and C. Huang. Interactive maps for a digital video library. *IEEE MultiMedia*, 7(1):60–67, 2000.
4. J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Commun.*, 37(1-2):89–108, 2002.
5. A. Hauptmann and M. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *Advances in Digital Libraries*, pages 168–179, 1998.
6. R. Kumar, H. Sawhney, J. Asmuth, A. Pope, and S. Hsu. Registration of video to geo-referenced imagery. In *Proc. of 14th Int'l Conf. on Pattern Recognition*, volume 2, pages 1393–1400, 1998.
7. T. Sato, T. Kanade, E. Hughes, M. Smith, and S. Satoh. Video OCR: indexing digital new libraries by recognition of superimposed captions. *Multimedia Syst.*, 7(5):385–395, 1999.
8. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of 9th IEEE Int'l Conf. on Computer Vision, Vol. 2*, 2003.
9. D. Sleator and D. Temperley. Parsing english with a link grammar. In *Third Int'l Workshop on Parsing Technologies*, 1993.
10. J. Yang and A. G. Hauptmann. Naming every individual in news video monologues. In *Proc. of the 12th ACM Intl. Conf. on Multimedia*, pages 580–587, 2004.