Using Closed Captions as Supervision for Video Activity Recognition

Sonal Gupta[*]
Department of Computer Science
Stanford University
353 Serra Mall
Stanford, CA 94305, USA
sonal@cs.stanford.edu

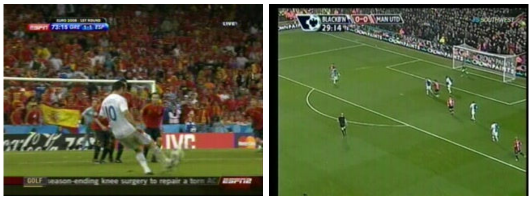
Raymond J. Mooney
Department of Computer Science
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712, USA
mooney@cs.utexas.edu

Abstract

Recognizing activities in real-world videos is a difficult problem exacerbated by background clutter, changes in camera angle & zoom, and rapid camera movements. Large corpora of labeled videos can be used to train automated activity recognition systems, but this requires expensive human labor and time. This paper explores how closed captions that naturally accompany many videos can act as weak supervision that allows automatically collecting 'labeled' data for activity recognition. We show that such an approach can improve activity retrieval in soccer videos. Our system requires no manual labeling of video clips and needs minimal human supervision. We also present a novel caption classifier that uses additional linguistic information to determine whether a specific comment refers to an ongoing activity. We demonstrate that combining linguistic analysis and automatically trained activity recognizers can significantly improve the precision of video retrieval.

1. Introduction

Due to the growing popularity of multimedia content, the need for automated video classification and retrieval systems is becoming increasingly important. In the past, video activity recognition and retrieval systems focused on datasets recorded in simplified settings that did not have much noise, e.g. the KTH (Schuldt, Laptev, and Caputo 2004) and Weizmann (Blank et al. 2005) datasets. Recently, significant progress has been made on activity recognition systems that detect specific human actions in real-world videos (Efros et al. 2003; Laptev et al. 2008). One application of recent interest is retrieving clips of particular events in sports videos such as baseball broadcasts (Fleischman and Roy 2007). Activity recognition in sports videos is particularly difficult because of background clutter, rapid change of actions, change in camera zoom and angle etc. Currently, the most effective techniques rely on supervised training data in the form of labeled video clips for each activity. Unfortunately, manually labeling videos is an expensive, time-consuming task.



(a) Kick: "Karagounis' free kick on to the head of no question, he had the job done before he slipped"

(b) Save: "I think Brown made a wonderful fingertip save there."

Figure 1: Examples of class 'kick' and 'save' along with their associated captions.

Broadcast and DVD videos increasingly include closed captions. Closed captions give a timestamped transcription of the audio portion of the program. This text can provide useful information about possible activities in videos "for free." To reduce human labor, one can exploit the weak supervisory information in captions such as sportscaster commentary. A number of researchers have proposed using closed captions or other linguistic information to enhance video retrieval, video classification, or speech recognition systems (see the related work section).

We propose a new approach that uses captions to automatically acquire "weakly" labeled clips for training a supervised activity recognizer. First, one selects keywords specifying the events to be detected. The system then finds these keywords (and their morphological variants) in captions of a video corpus and extracts video clips surrounding each retrieved caption. Sample captioned clips are shown in Figure 1. Although captions in sports video are useful clues about pictured activities, they are not definitive. Apart from the events in the game, sportscasters also talk about facts and events that do not directly refer to current activities. Therefore, the labeled data collected in this manner is very noisy. However, we show that there is enough signal in captions to train a useful activity recognizer. Although the accuracy of the weakly-trained recognizer is quite limited, it can be used to rerank caption-retrieved clips to present the most likely instances of the desired activity first. We present results on real soccer video showing that this approach can use video content to improve the precision of caption-based video retrieval without requiring *any* additional human su-

[*] Work done at the University of Texas at Austin

pervision. Our approach is scalable and can acquire a large amount of automatically labeled data given only a sizable corpus of captioned videos. Though we present our experiments on soccer games, the approach is generic since it does not use any domain-specific information.

As discussed before, captions are not definitive indicators of specific activities in a video. To further increase precision, we also present a method that uses a word-subsequence kernel (Bunescu and Mooney 2005; Lodhi et al. 2002) to classify captions according to whether or not they actually refer to a current event. The classifier learns phrases indicating a description of a current event versus an extraneous comment. Training this classifier requires some human labeling of captions; however this process is independent of the activities to be recognized and only needs to be done once for a given domain, such as sportscasting. To demonstrate this generality, we present experimental results showing *transfer learning* from soccer captions to baseball captions, when the classifier is trained on soccer captions and only a small number of baseball captions and tested on baseball captions. Transfer learning aims to improve accuracy in a target domain by using knowledge previously acquired in different but related source domains (Thrun and Pratt 1998).

Finally, we also show that combining the weakly-trained video classifier and the caption classifier improves the precision of activity retrieval more than either approach alone. Gupta (2009) discusses this work in detail.

2. Related Work

Activity recognition in videos has attracted significant attention in recent years (Efros et al. 2003; Schuldt, Laptev, and Caputo 2004). Activity classifiers are usually trained using human-labeled video clips. Recently, there has been increasing interest in using related text (such as closed captions, scripts, tags) and audio information together with visual information for various recognition and indexing tasks.

Recent work by Fleischman and Roy (2007) is the most closely related prior research. They used both captions and motion descriptions for baseball video to retrieve relevant clips given a textual query. They have also presented a method for using speech recognition on the soundtrack to further improve retrieval (Fleischman and Roy 2008). Unlike our approach, their system performs extensive video preprocessing to extract high-level, domain-specific video features, like “pitching scene” and “outfield”. Training these high-level feature extractors for preprocessing videos required collecting human-labeled video clips. Babaguchi, Kawai, and Kitahashi (2002) present an approach to event-based video indexing using collaborative processing of visual and closed-caption streams for sports videos. Their approach requires prior domain knowledge in the form of a hierarchy of events and manually-constructed keyword patterns used to identify events in the closed caption stream.

Several researchers have recently proposed systems that use information in associated scripts (available on the web) to improve classification, object detection, scene segmentation, and video retrieval (Everingham, Sivic, and Zisserman 2006; Laptev et al. 2008; Cour et al. 2008; Marszalek, Laptev, and Schmid 2009). Scripts provide detail descriptions of scenes and actions (such as *“John opens the car door”*), distinct from the spoken dialog. However, these methods cannot be used for domains such as sports videos that do not have associated scripts.

In contrast to this prior work, our approach uses words in captions as noisy labels for training a general-purpose, state-of-the-art, supervised activity recognizer without requiring *any* human labeling of video clips. In addition, our work does not need associated scripts, which are not available for most videos. We also present a novel caption classifier that classifies sentences in sports commentary as referring to a current event or not. This caption classifier is generic and independent of the activities to be detected and only requires humans to label a corpus of representative captions.

3. Approach

Figure 2 presents a diagram of our overall system. We first describe our procedure for automatically collecting labeled clips from captioned videos. We then explain the encoding of videos using motion descriptors and how they are used to train a video classifier. Next, we describe our caption classifier, and finally we explain the overall system for retrieving and ranking relevant clips.

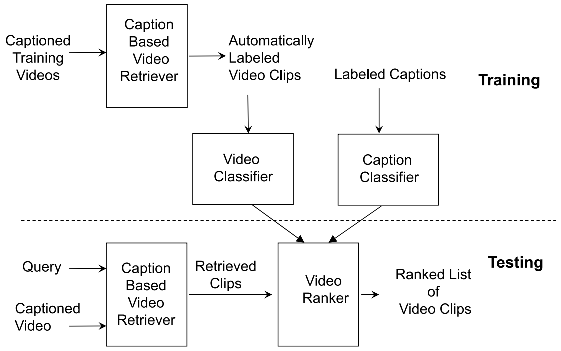

Captioned Training Videos
Caption Based Video Retriever
Automatically Labeled Video Clips
Labeled Captions
Training
Video Classifier
Caption Classifier
Query
Captioned Video
Caption Based Video Retriever
Retrieved Clips
Video Ranker
Ranked List of Video Clips
Testing

Figure 2: An overview of our video retrieval system.

3.1 Automatically Acquiring Labeled Clips

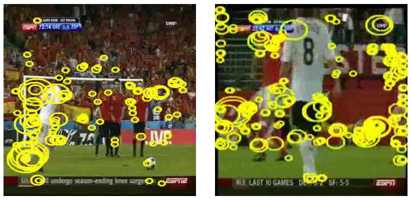
Videos, particularly sports broadcasts, generally have closed captions that can provide weak supervision about activities in the corresponding video. We use a simple method for extracting labeled video clips using such captions. However, captions in sports broadcasts are frequently broken into overlapping phrases. We first reconstruct full sentences from the stream of closed captions using a simple heuristic. Next, we identify all closed-caption sentences in a soccer game that contain exactly one member of a given set of activity keywords (currently, *save*, *kick*, *touch*, and *throw*). We also match alternative verb tenses, for example *save*, *saves*, *saved*, and *saving*. In our experiments, the number of potential clips that are rejected because their captions contained multiple query terms was about 2%, thus constraining the system to choose clips with exactly one keyword does not significantly affect the results. We then extract a fixed-length clip around the corresponding time in the video. In our dataset, we found that extracting 8-second clips was sufficient to capture the activities of interest. In live sports broadcasts, there is a significant lag between the video and the closed captions. We correct the correspondence between the caption timestamp and the video time to account for this lag. Each clip is then labeled with the corresponding keyword. For example, if the caption “What a nice kick!” occurs at time 00:30:00, we extract a clip from time 00:29:56 to 00:30:04 and label it as ‘kick’. Given a large corpus of captioned video, this approach can quickly assemble many labeled examples with no additional human assistance.

3.2 Motion Descriptors and Video Classification

Next, we extract visual features from each labeled video clip and represent it as a “bag of visual words.” We use features that describe both salient spatial changes and interesting movements. In order to capture non-constant movements that are interesting both spatially and temporally, we use the spatio-temporal motion descriptors developed by Laptev et al. (2008). These features have been shown to work well for human-activity recognition in real-world videos (Laptev and Perez 2007; Laptev et al. 2008; Marszalek, Laptev, and Schmid 2009). In addition, this approach can be used to detect activities in many domains since it does not use any domain-specific features or prior domain knowledge.

First, a set of interest points are extracted from a video clip. At each interest point, we extract a HoG (Histograms of oriented Gradients) feature and a HoF (Histograms of optical Flow) feature computed on the 3D video space-time volume. The patch is partitioned into a grid with 3x3x2 spatio-temporal blocks. Four-bin HoG and five-bin HoF descriptors are then computed for all blocks and concatenated into a 72-element and 90-element descriptors, respectively. We then concatenate these vectors to form a 162-element descriptor. A randomly sampled set of 117,000 motion descriptors from all video clips is then clustered using K-means(k=200) to form a vocabulary or “visual codebook”. Finally, a video clip is represented as a histogram over this vocabulary. The final “bag of visual words” representing a video clip consists of a vector of k values, where the i’th value represents the number of motion descriptors in the video that belong to the i’th cluster. Figure 3 shows some sample frames with detected motion features. As shown, most motion features are detected on interesting and useful patches picturing aspects of player activity. However, when the players are small in size and there is significant background clutter, many interest points are also detected in the background.

We then use the labeled clip descriptors to train an activity recognizer. The activity recognizer takes a video clip as input and classifies whether it belongs to the output action category. We tried several standard supervised classification methods from WEKA (Witten and Frank 2005), including SVMs and bagged decision trees. We obtained the best results with DECORATE, an ensemble algorithm that has been shown to perform well with small, noisy training sets (Melville and Mooney 2003; Melville et al. 2004). The high degree of noise in the automatically extracted supervision made DECORATE a particularly successful method. We use WEKA’s J48 decision trees as the base classifier for both DECORATE and bagging. We use an RBF kernel (γ=0.01) for SVMs since it performed the best out of several kernels that were tested. We learned a binary classifier for each activity class, considering the automatically labeled clips for that class as positive examples and clips that belong to other classes as negative examples.



(a) kick (b) save

Figure 3: Sample frames from two activity classes with detected motion features.

Sentence	Label
Beautiful pull-back.	1
Not only goals , but experience in the Germans’ favor but this is the semifinal.	0
That is a fairly good tackle.	1
I think I would have saved that myself.	0
Turkey can be well-pleased with the way they started.	0
They scored in the last kick of the game against the Czech Republic.	0
And Dempsey, with the first touch.	1
Mehmet Aur Elio, all it needed was a touch from Semih Senturk.	0
Cuba earns a corner kick.	1
Got kicked in the face.	0

Table 1: Some examples of captions with their labels in our dataset. Label ‘1’ means that the caption is relevant to some event in the game.

3.3 Identifying Relevant Captions

Sportscaster commentaries often include sentences that are not related to the current activities in the video. These sentences introduce noise in the automatically labeled video clips. For example, if a caption is “They really need to win this game to **save** their reputation.”, the algorithm will extract a clip corresponding to this sentence and label it as a ‘save’ instance, which is obviously a mistake. Therefore, we also train a caption classifier that determines whether or not a sentence actually refers to a current event in the video. When training this classifier, we use sample caption sentences manually labeled as relevant (1) or irrelevant (0). Examples of labeled captions are shown in Table 1. We expect the system to learn that phrases like ‘last game’, ‘needed touch’ are irrelevant to events going on in the video, and phrases like ‘earns kick’, ‘first touch’ are relevant.

A string subsequence kernel is particularly useful for learning such phrases, which otherwise are not captured by most commonly used bag-of-words text-classification methods. We use an SVM string classifier that uses a subsequence kernel (Lodhi et al. 2002), which measures the number of word subsequences shared by two strings. We use two subsequence patterns: word and Part-of-Speech (POS) subsequences. POS tags include potentially useful information such as verb tense. Such combined lexical and syntactic subsequence kernels have been shown to be useful for other natural-language processing tasks such as relation extraction (Bunescu and Mooney 2005). The Stanford POS tagger (Toutanova et al. 2003) was used to obtain POS tags for each word and we used LibSVM (Chang and Lin 2001) to learn a probabilistic caption classifier using this kernel.

Note that the caption classifier is trained once and is independent of the number or type of activities to be recognized since it only determines whether or not the sentence describes a current event. Also, humans labeled the captions in the training data without viewing the corresponding video. One might expect to need to look at the corresponding video when labeling captions; but as can be seen from the examples in Table 1, labeling captions alone is fairly unambiguous.

3.4 Retrieving and Ranking Videos

Given a new soccer game, our task is to retrieve video clips that contain a particular activity and present them in ranked order from most to least relevant. Given an activity keyword, we first retrieve videos using the captions alone as explained previously. As previously mentioned, we have considered four queries: *kick*, *save*, *throw* and *touch*. For each query i, a set of clips S_i are retrieved from the game. The goal is to rank the clips in S_i so that the truly relevant clips are higher in the ordered list of retrievals. The ranking is evaluated by comparing it to a correct human-labeling of the clips in S_i. Note that we use human-labeled video clips only to evaluate the quality of ranked retrievals.

One way to rank clips is to just use the automatically trained video classifier (called VIDEO). The video classifier assigns a probability $P(label|clip)$ to each retrieved clip according to the confidence it has that the clip belongs to the particular class, and the clips are ranked according to this probability. Another way to rank the clips is to just use the caption classifier (called CAPTION). The caption classifier assigns a probability $P(relevant|clip\text{-}caption)$ to each clip based on whether its corresponding caption is believed to describe an event currently occurring in the game. The classifier is expected to assign a higher probability to relevant clips. Since these two approaches use different information to determine relevance, we also aggregate their rankings using a linear combination of their probability assignments (called VIDEO+CAPTION):

$$\begin{aligned} P(label|clip\ with\ caption) &= \alpha P(label|clip) \\ &+ (1-\alpha)P(relevant|clip\text{-}caption) \end{aligned} \tag{1}$$

The value of α is determined empirically as described in Section 4.2.

4. Experiments

This section presents an experimental evaluation of our approach. First we describe the data we collected, next we explain our experimental methods, and finally we present the results.

Query Class	# Total	# Correct	% Noise
kick	303	120	60.39
save	80	47	41.25
throw	58	26	55.17
touch	183	122	33.33

Table 2: The number of total and correct clips for each category, along with the percentage of incorrect clips.

4.1 Dataset

Our primary dataset consists of 23 soccer games recorded from live telecasts. These games include corresponding time-stamped captions. Each game is around 1 hour and 50 minutes with an average of 1,246 caption sentences. The difficulty and diversity of the dataset can be seen from Figure 1. There is a wide difference in camera angle and zoom among the clips for a category, apart from activity occlusion and high background noise. Sometimes, the players are so small that even humans have difficulty in labeling the clips. We extracted clips for four activity keywords: $\{kick, save, throw, touch\}$. For evaluation purposes only, we manually labeled this data to determine the *correct* clips for each class, i.e. ones that actually depict the specified activity. The system itself never uses these gold-standard labels.

Table 2 shows the total number of clips for each keyword, as well as the number of correct clips and the amount of noise in each class (percentage of clips that are not correct). Note that the automatically labeled data extracted using captions is extremely noisy. We can see that the noise level is particularly high for 'kick' and 'throw.' The class 'kick' has the most noise, because in addition to irrelevant captions, the word 'kick' has two meanings in soccer commentary: kicking the ball, and kicking a person. We consider the former as the correct meaning for the query 'kick'.

The caption classifier was trained using a disjoint set of four games. Each sentence in the text commentary of these games was manually labeled as *relevant* or *irrelevant* to the current activity in the game. To reduce human time and effort, this labeling was performed without examining the corresponding video. All 4,368 labeled captions in this data, including 1,371 captions labeled as *relevant*, are used to train the final caption classifier.

4.2 Methodology

We performed experiments using a leave-one-game-out methodology, analogous to k-fold cross validation. In each fold, we left out one of the 23 games for testing and used the remaining 22 games for collecting automatically labeled data for training the video classifier. To select the value for α in Eq 1, in every fold, we randomly selected two games in the training set as a hold-out set and trained on the remaining games. We then selected the value of α that performed the best on the held-out portion of the training data and finally retrained on the full training set and tested on the test set.

For each query ($kick$, $save$, $throw$, $touch$), we retrieve and rank clips in the test game as explained in Section 3.4. We measure the quality of a ranking using Mean Average Precision (MAP), a common evaluation metric from information retrieval that averages precision (the percentage of retrieved items that are correct) across all levels of recall (the percentage of correct items that are retrieved) produced using different thresholds for a given set of ranked retrievals (Manning, Raghavan, and Schütze 2008).

We compare our approach to a simple baseline in which the clips retrieved using a given keyword are ranked randomly (called Baseline). We also compare our system to an idealized version in which the video classifier is trained using only the correct clips for each category (as determined by the human labeling used for evaluation) as positive examples (called Gold-Video).

4.3 Results

Ranking using the Video Classifier Table 3 shows MAP scores for ranking the clips using the video classifier trained using different learning methods. Video performs ~5 percentage points better than the baseline when Decorate is used, which is the best classifier due to its robustness to noisy training data. One interesting result is that, when using Decorate, Video even performs better than Gold-Video. For Bagging and SVM, Gold-Video performs better than Video, as expected. We suspect the reason why Video performs better when using Decorate is because the noise in the training examples actually helps build an even more diverse ensemble of classifiers, and thereby prevents over-fitting the gold-standard training examples. Video with SVM performs the worst. To avoid overfitting in SVM, we tried several values of the regularization parameter (C) and present the best results. Since bagging is also known to be fairly robust to noise, we suspect that SVM is overfitting the highly-noisy training data. In subsequent results, the video classifier was trained with Decorate since it gave the best performance.

Classifier	Decorate	Bagging	SVM
Baseline	65.68	65.68	65.68
Video	**70.75**	69.31	66.34
Gold-Video	67.8	70.5	67.20

Table 3: MAP scores when ranking using a video classifier.

Ranking using the Caption Classifier As explained in Section 3.4, the caption classifier can also be used to rank results. The MAP score for ranking with the caption classifier is shown in Table 4. Caption performs ~5 percentage points better than the baseline, demonstrating the value of using linguistic information to decide whether or not a caption describes an ongoing event. It is interesting to note that Video and Caption perform almost the same, even though the former exploits visual information while the latter uses linguistic cues.

Caption Classifier Accuracy The caption classifier also performs reasonably well on its specific task of identifying relevant captions. This was evaluated using leave-one-game-out on the four games used to train the final caption classifier. An SVM with a subsequence kernel that includes both word and POS subsequences (word+pos ssk) was 79.81% accurate, compared to 79.26% for a subsequence kernel using just word subsequences (word ssk), and to a baseline of 69.02% when captions are labeled with the majority class. An SVM with standard bag-of-words features gave an accuracy of only 69.07%, barely beating the 69.02% baseline of guessing the majority class (negative), thereby verifying the importance of using word order and subsequences when classifying short caption sentences. In our video retrieval experiments, we used a word+pos subsequence kernel, although we would expect similar results with just a word subsequence kernel.

Approach	MAP
Baseline	65.68
Caption	70.747
Video	70.749
Video+Caption	**72.11**
Gold-Video+Caption	70.53

Table 4: MAP scores for different ranking methods.

Approach	Training Dataset	Accuracy
Majority Class Baseline	Soccer and Baseball	69.23
Bag-of-Words	Soccer and Baseball	66.39
word ssk	Soccer and Baseball	72.07
word+pos ssk	Soccer and Baseball	66.69
word ssk	Soccer	71.97
word ssk	Baseball	67.59

Table 5: Classification accuracy of the caption classifier when tested on baseball captions.

To demonstrate the generality of the caption classifier across different sports domains, we also used it to classify baseball captions (relevant vs. irrelevant) when trained on soccer captions and (optionally) a small number of baseball captions. The content of these two sports commentaries are quite different since sportscasters tend to use many game-specific words and phrases. The baseball data consists of 985 captions from a baseball game manually labeled as ‘relevant’ or ‘irrelevant’. Table 5 shows results of *transfer learning* from soccer to baseball captions. Transfer learning is generally performed using a large out-of-domain (source) training set and a small or empty in-domain (target) training set. The small baseball dataset was split into training and test using five-fold cross validation. Therefore, the training set consisted of all of the labeled soccer captions (4,368 examples) and 80% of the labeled baseball captions (788 examples). In the table, ‘Soccer and Baseball’ means both soccer and baseball captions were used for training. ‘Soccer’ means only soccer captions (source data) were used for training, and similarly ‘Baseball’ means only baseball captions (target data) were used. word ssk using both Soccer and Baseball training data performs the best, verifying the importance of using both word order and transfer learning. However, using POS tags (WORD+POS SSK), actually hurt performance. WORD SSK using only Baseball training data performs even worse than the baseline because there are not enough training examples. It is interesting to note that WORD SSK trained only on Soccer data performs better than most other methods, even though the training set does not contain even a single example from the baseball domain.

Aggregating the rankings The video and caption classifiers leverage two different sources of information, visual and linguistic, respectively. Table 4 demonstrates that combining the two sources of information (VIDEO and CAPTION) increases the MAP score another ~1.5 percentage points over the individual classifiers and ~6.5 percentage points over the baseline. All results in Table 4 are statistically significant compared to the BASELINE according to a one-tailed paired t-test with a 95% confidence level. The average value of α, computed using internal cross-validation in each trial (see Section 4.2), was 0.46.

5. Future Work

Exploiting the multi-modal nature of captioned videos is a vast and little-explored area, and there are many areas ripe for further investigation. Improving the supervised activity recognizer is a major area for future research. A potentially promising approach is to preprocess the video to remove background clutter and focus on the activity of the players on the field.

We have shown that our approach improves the *precision* of a caption-based video retrieval system by reranking clips that were retrieved using the captions alone. Improving *recall* would require scanning the entire video with a trained activity recognizer in order to extract additional clips that are *not* accompanied by the corresponding activity keyword. This is a computationally expensive process and evaluating recall requires manual labeling of the entire video; therefore, this was left for future research.

Another promising direction is to exploit temporal relations between activities to improve the final classifier as well as help collect more labeled data. For example, the probability of a video clip being a 'save' should be higher if we know that the immediately preceding clip is a 'kick'. Hidden Markov Models and Conditional Random Fields could potentially be used to model such temporal relations.

6. Conclusion

This paper has demonstrated that closed captions can be used to automatically train a video activity recognizer without requiring *any* manual labeling of video clips, and that this activity recognizer can be used to improve the precision of caption-based video retrieval. In addition, we showed that training a classifier to identify captions that describe current activities can improve precision even further. Additional results demonstrated that such a caption classifier trained for soccer also generalizes to other sports domains like baseball. Finally, the encouraging results on aggregating retrieval rankings from both video and caption classifiers provides additional evidence that exploiting the multimodal nature of closed-captioned video can improve the effectiveness of activity recognition and video retrieval.

Acknowledgements This research author was funded by NSF grant IIS-0712907. We would like to thank Ruchica Behl for labeling a part of the soccer captions. We are also thankful to Kristen Grauman and the anonymous reviewers for their valuable comments.

References

Babaguchi, N.; Kawai, Y.; and Kitahashi, T. 2002. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*.

Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; and Basri, R. 2005. Actions as space-time shapes. In *ICCV 2005*.

Bunescu, R. C., and Mooney, R. J. 2005. Subsequence kernels for relation extraction. In *NIPS 2005*.

Chang, C.-C., and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*.

Cour, T.; Jordan, C.; Miltsakaki, E.; and Taskar, B. 2008. Movie/script: Alignment and parsing of video and text transcription. In *ECCV 2008*.

Efros, A. A.; Berg, A. C.; Mori, G.; and Malik, J. 2003. Recognizing action at a distance. In *ICCV 2003*.

Everingham, M.; Sivic, J.; and Zisserman, A. 2006. Hello! My name is... Buffy – Automatic naming of characters in TV video. In *BMVC 2006*.

Fleischman, M., and Roy, D. 2007. Unsupervised content-based indexing for sports video retrieval. In *Workshop on Multimedia Information Retrieval 2007*.

Fleischman, M., and Roy, D. 2008. Grounded language modeling for automatic speech recognition of sports video. In *ACL 2008: HLT*. Association for Computational Linguistics.

Gupta, S. 2009. Activity retrieval in closed captioned videos. Master's thesis, University of Texas at Austin.

Laptev, I., and Perez, P. 2007. Retrieving actions in movies. In *ICCV 2007*.

Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *CVPR 2008*.

Lodhi, H.; Saunders, C.; Shawe-Taylor, J.; Cristianini, N.; and Watkins, C. 2002. Text classification using string kernels. *JMLR*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Marszalek, M.; Laptev, I.; and Schmid, C. 2009. Actions in context. In *CVPR 2009*.

Melville, P., and Mooney, R. J. 2003. Constructing diverse classifier ensembles using artificial training examples. In *IJCAI 2003*.

Melville, P.; Shah, N.; Mihalkova, L.; and Mooney, R. J. 2004. Experiments on ensembles with missing and noisy data. In *Workshop on Multi Classifier Systems*.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local SVM approach. In *ICPR 2004*.

Thrun, S., and Pratt, L. 1998. Learning to learn: Introduction and overview. In *Learning To Learn*. Kluwer Academic Publishers.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL 2003*.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (2nd edition)*. Morgan Kaufman Publishers.