

Using Closed Captions to Train Activity Recognizers that Improve Video Retrieval

Sonal Gupta and Raymond J. Mooney

Department of Computer Sciences

The University of Texas at Austin

1 University Station C0500

Austin, Texas, USA 78712-0233

{sonaluta,mooney}@cs.utexas.edu

Abstract

Recognizing activities in real-world videos is a difficult problem exacerbated by background clutter, changes in camera angle & zoom, rapid camera movements etc. 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. Recently, significant progress has been made on activity recognition systems that detect specific human actions in real-world videos [6, 15]. One application of recent interest is retrieving clips of particular events in sports videos such as baseball broadcasts [9]. Activity recognition in sports videos is particularly difficult because of the ambiguous video cues, background clutter, rapid change of actions, change in camera zoom and angle etc. Currently, the most effective techniques for activity recognition rely on supervised training data in the form of labeled video clips for particular classes of actions. Unfortunately, manually labeling videos is an



Figure 1. Examples of class ‘kick’, ‘save’, ‘throw’, and ‘touch’ along with their associated captions.

expensive, time-consuming task.

As an alternative, closed captions can provide useful information about possible activities in videos for “free.” Closed captions are increasingly available for most broadcast and DVD videos. 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 sound recognition systems [8, 10, 1, 15, 5] (see Section 2). We propose a new approach that uses captions to automat-

ically acquire “weakly” labeled clips for training a supervised activity recognizer. First, one selects keywords specifying the events to be detected. As an example, we present results for four activity keywords for soccer videos: *kick*, *save*, *throw* and *touch*. Sample captioned clips are shown in Figure 1. The system then finds these keywords (and their morphological variants) in captions of a relevant video corpus and extracts video clips surrounding each retrieved caption. Although captions in sports video are useful clues about activities in video, 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. For example, a sportscaster might say ‘*He scored a great goal in the last game*’. 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 the 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 supervision.

To further increase precision, we also propose using a word-subsequence kernel [16, 3] to classify captions as to whether or not they actually refer to a current event. The classifier learns subsequences of words 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. Our results show that using this caption classifier to rerank retrieved clips to prefer those commenting on a current event also improves precision. Finally, we also show that combining the weakly-trained video classifier and the caption classifier improves precision more than either approach alone. A pictorial overview of the complete system is shown in Figure 2.

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 presents our approach, Section 4 describes our experimental methodology and results, and Sections 5 and 6 present future work and conclusions.

2. Background and Related Work

Activity recognition in videos has attracted significant attention in recent years. Many researchers have developed activity recognizers using only visual cues and hand-labeled video clips [21, 6, 23, 14, 2]. There has also been increasing interest in using textual information along with visual information for various tasks. Nitta *et al.* [19] annotated sports video by associating text segments with image segments. Their approach uses prior knowledge of the

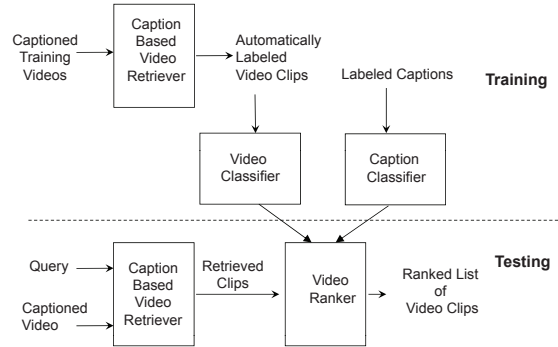


Figure 2. An overview of our video retrieval system

game and the key phrases generally used in its commentary. Gupta *et al.* [12] used captions and visual information in sports video as two views for semi-supervised classification with co-training. Ozkan and Duygulu [20] associated news videos with words to perform scene and object recognition, but used keyframes for recognition and thus did not use motion cues. Everingham *et al.* [7], Laptev *et al.* [15], Cour *et al.* [5] incorporated visual information, closed-captioned text, and movie scripts (with scene descriptions) to automatically annotate videos in movies and then use them for classification, retrieval and annotation of videos. Their methods cannot be used for domains such as sports videos that do not have associated scripts. Laptev *et al.* [15] used captions of labeled clips to learn a text classifier to identify whether the text corresponding to a clip is representative of the clip activity. Then, using a set of extracted representative clips, they trained a video classifier to classify human actions.

Recent work by Fleischman and Roy is the most closely related prior research. Fleischman and Roy [8] used both captions and motion descriptions for baseball video to retrieve relevant clips given a textual query. Additionally, Fleischman and Roy [9] presented a method for using speech recognition on the soundtrack to further improve retrieval. They used an unsupervised Author Topic Model, a generalization of Latent Dirichlet Allocation, to learn correlations between caption text and encoded event representations. Unlike our approach, their system performed extensive video preprocessing to extract high-level, domain-specific video features, like “pitching scene” and “outfield.” Training these high-level feature extractors required collecting human-labeled video clips.

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 a rich source of description of events in a video but not available for most of the 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

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

3.1. Automatically Acquiring Labeled Data

Videos, particularly sports broadcasts, generally have closed captions that provide weak supervision about activities in the corresponding video. We use a simple method for extracting labeled video clips using these captions. Captions in sports broadcasts are frequently broken into overlapping phrases. We first use a simple heuristic method to reconstruct full sentences from the stream of closed captions. 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 matched alternative verb tenses, for example *save*, *saves*, *saved*, and *saving*. We then extract a fixed-length clip (currently 8 seconds) around the corresponding time in the video. 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’. The algorithm for acquiring labeled clips could be made more sophisticated by exploiting additional linguistic and visual information, but our results demonstrate that even this simple approach suffices to obtain useful results. 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 “bags 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

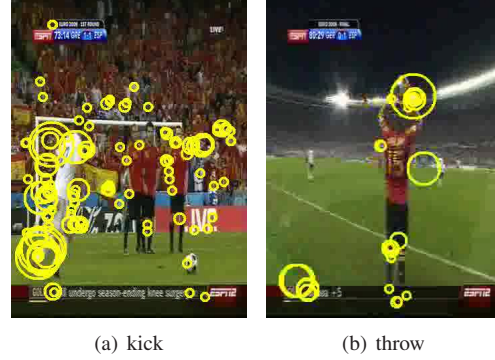


Figure 3. Example frames from two query classes with detected motion features

use the spatio-temporal motion descriptors developed by Laptev *et al.* [15]. We chose the spatio-temporal interest point approach over a dense optical flow-based approach in order to provide a scale-invariant, compact representation of activity in the scene.

To detect spatio-temporal events, Laptev *et al.* [15] builds on Harris and Forstner’s interest point operators [13, 11] and detects local structures where the image values have significant local variation in both space and time.

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 the motion descriptors from all video clips is then clustered to form a vocabulary or “visual codebook”. We use K-means ($k=200$) with 117,000 feature vectors sampled randomly from the corpus of clips. 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 3.2 shows example frames of query class ‘kick’ and ‘throw’ with detected motion features.

We then use the labeled clip descriptors to train an activity recognizer. We tried several standard supervised classification methods from WEKA [24], including SVMs and bagged decision trees. However, we obtained the highest accuracy with DECORATE, an ensemble algorithm that has been shown to perform well with small, noisy training sets [17, 18]. 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 build a

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

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.

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. We also tried one-against-one classifiers, but they gave inferior performance.

The approach can be made scalable in terms of number of queries by clustering the queries and then representing every cluster as a class.

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 one of the captions 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’, 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 the classifier, we use sample caption sentences manually labeled as relevant (1) or irrelevant (0). Examples of labeled captions are shown in Table 1.

We use an SVM string classifier that uses a subsequence kernel [16], which measures how many subsequences are shared by two strings. A subsequence is any ordered sequence of tokens occurring either contiguously or non-contiguously in a string. By using word order, a subsequence kernel can exploit syntactic cues unavailable to a standard “bag of words” text classifier; therefore, we found that it obtained superior accuracy for determining caption relevance. Bunescu and Mooney [3] proposed a generalization of subsequence kernels that integrates information from multiple subsequence patterns. We use two subsequence patterns: word subsequences and Part-of-Speech (POS) subsequences. The Stanford POS tagger [22] was used to obtain POS tags for each word and we used LibSVM [4] 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. Also, humans labeled the captions in the training data without viewing the corresponding video. This may

introduce some noisy supervision but avoids the additional human burden of watching the video.

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 in Section 3.1. 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 to each retrieved clip ($P(\text{label}|\text{clip})$), 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 ($P(\text{relevant}|\text{clip-caption})$) to each clip based on whether its corresponding caption is believed to describe an event currently occurring in the game. The classifier should 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-and-CAPTION):

$$P(\text{label}|\text{clip with caption}) = \alpha P(\text{label}|\text{clip}) + (1 - \alpha) P(\text{relevant}|\text{clip-caption}) \quad (1)$$

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

4. Experiments

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. We extracted clips for four activity keywords: $\{\textit{kick}, \textit{save}, \textit{throw}, \textit{touch}\}$, as discussed in Section 3. 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.

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.

A disjoint set of four games was used to train the caption classifier. 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 were used to train the caption classifier. The dataset consists of 1,371 captions labeled as *relevant*.

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 Equation 1, in every fold, we randomly selected two games in the training set as a held 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.

We consider four queries for video retrieval: $\{kick, save, throw, touch\}$. For each query, 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 across all levels of recall for a given set of ranked retrievals. If the set of retrieved clips for a query $q_i \in Q$ is $\{clip_1, clip_2, \dots, clip_{m_i}\}$ and L_{ik} is the subset of the k highest-ranked clips, then

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{m_i} \sum_{k=1}^{m_i} Precision(L_{ik})$$

where $Precision(L_{ik})$ is defined as ratio of the number of *correct* clips in L_{ik} over the total number of clips in L_{ik} . We compare our approach to a simple baseline in which the clips 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 (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 advantage for noisy training data (see Section 3.2). 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 in the data. VIDEO with SVM performs the worst. We tried several values of the regularization parameter (C) in SVM 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 the results below, we assume the video classifier is trained with DECORATE since it performs the best. The actual accuracy of the learned classifiers is not that high, with an macro-average F-measure of 20%; however, they are still useful at improving the ranking of clips within each class.

Classifier	DECORATE	Bagging	SVM
BASELINE	65.68	65.68	65.68
VIDEO	70.749	69.31	66.34
GOLD-VIDEO	67.8	70.5	67.20

Table 3. MAP scores when ranking the retrieved clips 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 knowledge to decide whether or not a caption describes an ongoing event. The caption classifier performs reasonably well on the classification task as well. The classification methodology was leave-one-game-out on the four games that were used to build the final caption classifier. The classification accuracy of an SVM with a subsequence kernel that includes word and POS subsequences is 79.81%, compared to a baseline of 69.02% when all captions are labeled with the most frequent class. Using an SVM with a bag-of-words approach gave worse results than the baseline, signifying the importance of word order.

Approach	MAP
BASELINE	65.68
CAPTION	70.747
VIDEO	70.749
VIDEO+CAPTION	72.11
GOLD-VIDEO+CAPTION	70.53

Table 4. MAP measures for different approaches

Aggregating the rankings

The rankings of the video and caption classifiers leverage two different sources of information, visual and linguistic, respectively. Table 4 shows 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 as compared to BASELINE on a one-tailed paired t-test with a 95% confidence level. Table 5 shows MAP score for the four query classes for different approaches. Sometimes there are no correct instances of a query class in a game and the corresponding MAP score becomes *NaN*. Note that as we ignore *NaN* values while averaging MAP scores in leave-one-game-out cross-validation, the final MAP score is numerically not equal to the average of the MAP scores of query classes. We can see that VIDEO+CAPTION improves the MAP score most for the query class ‘touch’ and least for ‘kick’. This was expected as noise in the automatically labeled dataset was highest for ‘kick’ and lowest for ‘touch’ (see Table 2).

Table 6 show rankings from most to least relevant and the MAP scores computed by VIDEO, CAPTION, and VIDEO+CAPTION for the query class ‘touch’ for a test game. There were seven clips extracted from the game for the query class. The MAP score of VIDEO+CAPTION is higher than VIDEO and CAPTION individually. We can see that VIDEO and CAPTION classifiers leverage different information and aggregating them gives better results. For example, even though VIDEO ranks Clip2 and Clip4 higher, CAPTION gives them low rankings thus decreasing their rankings in VIDEO+CAPTION. Similarly, Clip7 was ranked high by CAPTION but VIDEO gives it a low ranking, pushing its ranking down when aggregating both rankings. Clip7, corresponding to the caption ‘lovely touch’, is not relevant to the class ‘touch’ as commentators were discussing an event that happened several seconds back and the video clip does not capture the event. Similarly, Clip2 is not relevant to the query class.

5. Future Work

Exploiting the multi-modal character of captioned videos is a vast and little-explored area, and there are many

Approach	kick	save	throw	touch
VIDEO+CAPTION	46.42	73.42	77.38	86.52
GOLD-VIDEO+CAPTION	46.7	75.57	76.27	82.56
BASELINE	46.13	69.33	72.97	75.77

Table 5. MAP score for every query class when different approaches are used

areas ripe for further investigation. Improving the supervised activity recognizer is a major area for future research. A promising approach is to preprocess the video to remove background clutter and focus on the activity of the players on the field. By focusing the activity recognizer on player actions, we believe accuracy could be significantly improved.

Since our best video classifier that is trained using noisy caption-based labeling already out-performs one trained on gold-standard data, it is not surprising that we found no improvement when using the video and/or caption classifier to automatically “clean” the caption-labeled data prior to training. However, given a better activity recognizer, we believe that using linguistic and video analysis to remove some of the false positives from the training data would further improve the results.

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. On the other hand, 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. Unfortunately, this is a very computationally expensive process, and properly evaluating recall would require the laborious task of manually labeling all of the relevant events in the entire video. Therefore, we have left this aspect of the evaluation to future research.

Our intuition is that exploiting temporal relations between activities will improve the video classifier as well as help collect more labeled data. For example, the probability of a video clip being of query class ‘save’ should be higher if we know that the clip preceding it in time is a ‘kick’.

Finally, it would be interesting to train the caption classifier on captions from one sport and test it on captions from another sport. Since the caption classifier is trying to detect a very abstract linguistic property (depiction of a current event) it should generalize fairly well to other domains.

6. Conclusion

In this paper, we have shown that closed captions can be used to automatically train an video activity recognizer without requiring *any* manual labeling of video clips. We have also demonstrated that this activity recognizer can be used to improve the precision of caption-based video retrieval. Finally, we have shown that training a caption clas-

sifier to identify captions that describe current activities can improve precision even further. This is further indication that exploiting the multimodal nature of closed-captioned video can improve the effectiveness of activity recognition and video retrieval technology.

Acknowledgment

We thank Ruchica Behl for valuable conversations and labeling a part of the caption data. We also thank Kristen Grauman and Tuyen N. Huynh for their useful suggestions. This work was funded by grant IIS-0712907X from the U.S. National Science Foundation. Most of the experiments were run on the Mastodon Cluster, provided by NSF Grant EIA-0303609.

References

- [1] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 2002.
- [2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision*, 2005.
- [3] R. C. Bunescu and R. J. Mooney. Subsequence kernels for relation extraction. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, Vancouver, BC, 2005.
- [4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [5] T. Cour, C. Jordan, E. Mitsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, 2008.
- [6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, Nice, France, 2003.
- [7] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy – Automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006.
- [8] M. Fleischman and D. Roy. Situated models of meaning for sports video retrieval. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, April 2007.
- [9] M. Fleischman and D. Roy. Unsupervised content-based indexing for sports video retrieval. In *Ninth ACM Workshop on Multimedia Information Retrieval (MIR)*, Augsburg, Germany, 2007.
- [10] M. Fleischman and D. Roy. Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT*, June 2008.
- [11] W. Forstner and E. Gulch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, 1987.
- [12] S. Gupta, J. Kim, K. Grauman, and R. J. Mooney. Watch, listen & learn: Co-training on captioned images and videos. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008)*, Antwerp, Belgium, September 2008.
- [13] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.
- [14] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, October 2007.
- [15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision and Pattern Recognition*, 2008.
- [16] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.
- [17] P. Melville and R. J. Mooney. Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-2003)*, pages 505–510, Acapulco, Mexico, August 2003.
- [18] P. Melville, N. Shah, L. Mihalkova, and R. J. Mooney. Experiments on ensembles with missing and noisy data. In *Proceedings of the Fifth International Workshop on Multi Classifier Systems (MCS-2004)*, Cagliari, Italy, 2004.
- [19] N. Nitta, N. Babaguchi, and T. Kitahashi. Extracting actors, actions and events from sports video - a fundamental approach to story tracking. In *ICPR '00: Proceedings of the International Conference on Pattern Recognition*, Washington, DC, USA, 2000. IEEE Computer Society.
- [20] D. Ozkan and P. Duygulu. Finding people frequently appearing in news. In *CIVR*, pages 173–182, 2006.
- [21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04)*, Washington, DC, USA, 2004.
- [22] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, 2003.
- [23] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. In *2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation*, 2007.
- [24] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (2nd edition)*. Morgan Kaufman Publishers, San Francisco, 2005.















Ranking using VIDEO MAP score = 73.33	Ranking using CAPTION MAP score = 67.91	Ranking using VIDEO+CAPTION MAP score = 80.41
<p>✓Clip1</p>  <p>↓</p> <p>✗Clip2</p>  <p>↓</p> <p>✓Clip3</p>  <p>↓</p> <p>✗Clip4</p>  <p>↓</p> <p>✓Clip5</p>  <p>↓</p> <p>✓Clip6</p>  <p>↓</p> <p>✗Clip7</p> 	<p>✗Clip7: Lovely touch.</p> <p>↓</p> <p>✓Clip1: Just trying to touch it on.</p> <p>↓</p> <p>✓Clip3: When he comes back on the ball, just about got a touch on it just about.</p> <p>↓</p> <p>✓Clip6: Just touched on by Nani.</p> <p>↓</p> <p>✓Clip5: And the lovely little touch from Ryan Giggs.</p> <p>↓</p> <p>✗Clip2: If he had not touched it.</p> <p>↓</p> <p>✗Clip4: I do not think it was touched.</p>	<p>✓Clip1: Just trying to touch it on.</p>  <p>↓</p> <p>✗Clip7: Lovely touch.</p>  <p>↓</p> <p>✓Clip3: When he comes back on the ball, just about got a touch on it just about</p>  <p>↓</p> <p>✓Clip6: Just touched on by Nani.</p>  <p>↓</p> <p>✓Clip5: And the lovely little touch from Ryan Giggs.</p>  <p>↓</p> <p>✗Clip2: If he had not touched it.</p>  <p>↓</p> <p>✗Clip4: I do not think it was touched.</p> 

Table 6. Rankings, from most relevant to least relevant, using VIDEO, CAPTION and VIDEO+CAPTION for class 'touch' and the respective MAP scores for the query, for a test game. A check mark means according to the ground-truth labels, the clip is relevant to the query class and a cross mark means it is not.