Activity Retrieval in Closed Captioned Videos

APPROVED BY

SUPERVISING COMMITTEE:

____

Raymond J. Mooney, Supervisor

____

Kristen Grauman

Activity Retrieval in Closed Captioned Videos

by

Sonal Gupta, B.Tech.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF ARTS

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2009

Acknowledgments

I am thankful to a lot of people who helped me, directly and indirectly, to finish my Master's thesis. My adviser, Ray Mooney, deserves the most credit for believing in me and motivating me to work on the topic of my interest at all times. I will never forget his wonderful and candid advises for all paths of life. Apart from Ray, I am also very thankful to Kristen Grauman, who led me to pursue research in computer vision. She has been a great teacher and wonderful guide.

I am grateful to my family for their love, motivation and support throughout my life. My parents, Sudha and Arvind Gupta, supported my every decision and have been the best parents anyone would hope for. My sister Anshu has been my inspiration as I followed her into the field of Computer Science. I have special place in my heart for my grandfather Kashi Ram Gupta and my younger brother Ankur.

I am thankful to my wonderful friends at Austin. I found a great friend in my flatmate Pooja. My friends Karthik, Harshdeep, Alok, Pavithra, Amitanand, Ruchica, Guneet made my stay at Austin filled with unforgettable memories. I am especially thankful to Lily Mihalkova as she has been a great friend over the last year. She gave some remarkable suggestions to improve my thesis work. I would definitely miss her presence in my new office. I am grateful to Ruchica for labeling a part of the caption data. I would also like to thank Joohyun for his collaboration with me that led to me working with Ray. I want to thank the CS department and staff members who have always been very kind and helpful to me.

I appreciate the efforts of the UT Machine Learning group, especially Joohyun, Tuyen, David and Lily, for the insightful conversations and constructive comments on my papers and practice talks.

I am grateful to the National Science Foundation for funding this research through the grant IIS-0712907.

I finally want to thank Apurva Samudra for being beside me all the time. I could not have done without you.

Activity Retrieval in Closed Captioned Videos

Sonal Gupta, M.A.
The University of Texas at Austin, 2009

Supervisor: Raymond J. Mooney

Recognizing activities in real-world videos is a difficult problem exacerbated by background clutter, changes in camera angle & zoom, occlusion 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 thesis 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.

Table of Contents

List of Tables

List of Figures

Chapter 1

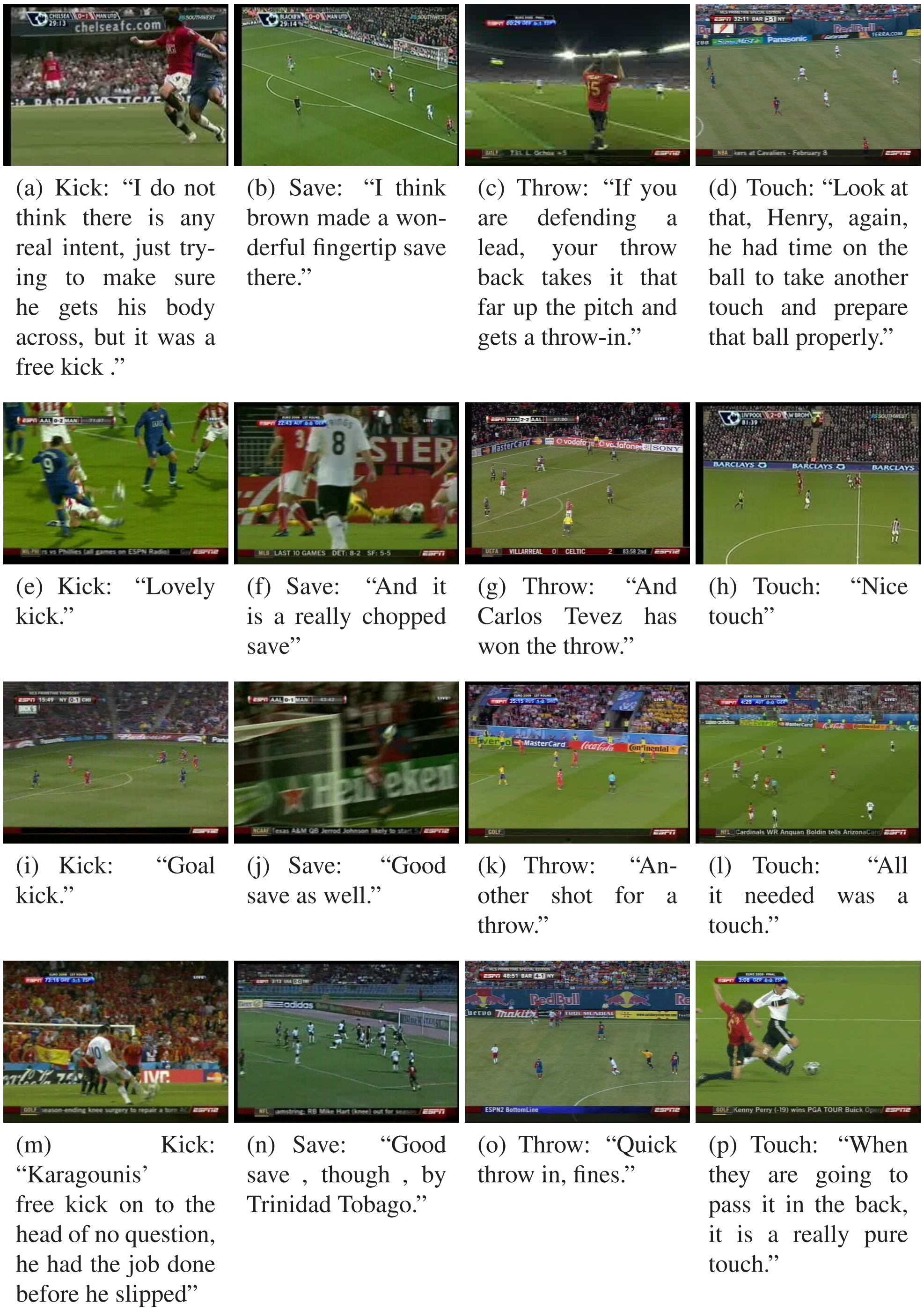
Introduction

Due to the growing popularity of multimedia content, the need for automated video classification and retrieval systems is becoming increasingly important. Annotating and indexing videos will be crucial for managing the worlds ever-growing creation of digital videos. Video classification and retrieval systems have wide practical use, such as in surveillance, video search and digital libraries. A video classifier classifies a video clip whether it belongs to a pre-specified set of categories. Video retrieval systems, on the other hand, are required to extract relevant clips from a video and rank them according to their relevance to either a text or video query. Video classification and retrieval often require activity or action recognition. Activity recognition is very hard because camera motion and zoom along with well-known static image recognition problems, such as illumination, occlusion, view point difference, make visual cues extremely ambiguous. In the past, video activity recognition and retrieval systems focussed on datasets recorded in simplified settings that did not have much noise (for e.g. KTH (Schuldt, Laptev, & Caputo, 2004) and Weizmann (Blank, Gorelick, Shechtman, Irani, & Basri, 2005) datasets). Recently, significant progress has been made on activity recognition systems that detect specific human actions in real-world videos (Efros, Berg, Mori, & Malik, 2003; Laptev, Marszalek, Schmid, & Rozenfeld, 2008). One application of recent interest is retrieving clips of particular events in sports videos such as baseball broadcasts (Fleischman & Roy, 2007b). Activity recognition in sports videos is particularly difficult because the settings in which the videos are recorded are less structured and there is rapid change in view point, zoom and angle. 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 expensive, time-consuming task.

As an alternative, broadcast and DVD videos increasingly have closed captions. Closed captions are timestamped transcription of the audio portion of the program [1]. These closed captions 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 sound recognition systems (Babaguchi, Kawai, & Kitahashi, 2002; Cour, Jordan, Miltsakaki, & Taskar, 2008; Fleischman & Roy, 2007a, 2007b, 2008; Laptev et al., 2008) (see Chapter 2).

We propose a new approach that uses captions to automatically acquire "weakly" labeled clips for training a supervised activity recognizer. Our approach is quite scalable in acquiring a large amount of automatically labeled data given a large corpus of captioned videos. First, one selects keywords specifying the events to be detected. As an example, we present results for four activity keywords for

[1] `http://en.wikipedia.org/wiki/Closed_captioning`



(a) Kick: “I do not think there is any real intent, just trying to make sure he gets his body across, but it was a free kick .”

(b) Save: “I think brown made a wonderful fingertip save there.”

(c) Throw: “If you are defending a lead, your throw back takes it that far up the pitch and gets a throw-in.”

(d) Touch: “Look at that, Henry, again, he had time on the ball to take another touch and prepare that ball properly.”

(e) Kick: “Lovely kick.”

(f) Save: “And it is a really chopped save”

(g) Throw: “And Carlos Tevez has won the throw.”

(h) Touch: “Nice touch”

(i) Kick: “Goal kick.”

(j) Save: “Good save as well.”

(k) Throw: “Another shot for a throw.”

(l) Touch: “All it needed was a touch.”

(m) Kick: “Karagounis’ free kick on to the head of no question, he had the job done before he slipped”

(n) Save: “Good save , though , by Trinidad Tobago.”

(o) Throw: “Quick throw in, fines.”

(p) Touch: “When they are going to pass it in the back, it is a really pure touch.”

Figure 1.1: Examples of class ‘kick’, ‘save’, ‘throw’, and ‘touch’ along with their associated captions.

soccer videos: *kick*, *save*, *throw* and *touch*. Sample captioned clips are shown in Figure 1.1. The system then finds these keywords (and their morphological variants) in captions of a 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. Though we present our experiments on soccer games, we believe the approach is generic as it does not use previous knowledge of the game, such as structure of the soccer game.

To further increase precision, we also propose using a word-subsequence kernel (Bunescu & Mooney, 2005; Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002) 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. To show this generality, we present experimental results showing *transfer learning* from soccer captions to baseball captions, when the classifier is trained on soccer captions and a part of baseball captions and tested on rest of the baseball captions. Transfer learning aims to improve accuracy on a target domain by using knowledge acquired while learning on the source domains (Thrun & Pratt, 1998).

The caption classifier ranks captions based on its prediction whether or not they refer to an event. Our results on video retrieval 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.

Our main contribution is a system to retrieve and 'weakly' label video clips using closed captions, and to integrate a video activity classifier trained using the weakly labeled video clips with a novel caption classifier. Earlier approaches have focussed on either activity recognition using manually labeled video clips, or acquiring the labeled data using scripts associated with videos. Scripts are detailed description of scenes and actions, in addition to the dialogues in the video. An example of script text is shown in Table 1.1 (see Laptev et al., 2008). As we can see, the description 'Rick sits down with Ilsa' is a very strong cue for a sitting action. Scripts, however, are not available for most of the videos, for example, sports videos. We show in this thesis that the closed captions associated with such videos provide enough information about the activities, and can be effectively used

Rick: Why weren't you honest with me? Why did you keep your marriage a secret?
01:20:17 - 01:20:23 Rick sits down with Ilsa.
Ilsa: Oh, it wasn't my secret, Richard. Victor wanted it that way.
Not even our closest friends knew about our marriage.

Table 1.1: An example of dialogues and detailed description of actions in a script. Script, though a valuable cue, is not available for most of the videos, for example, sports videos.

to build a useful retrieval system. Our approach can acquire weakly labeled video clips without using scripts, and we present out results on closed captioned soccer videos.

A pictorial overview of the complete system is shown in Figure 1.2. First, videos clips are retrieved and automatically labeled using the closed captions. We then build a video classifier using the labeled set. A caption classifier is separately built using labeled closed captions. During testing, given a query and a video, we retrieve clips using the closed captions and rank them using the video classifier and the caption classifier.

The rest of the thesis is organized as follows: Chapter 2 discusses related work, Chapter 3 provides some background needed in remainder of the thesis, Chapter 4 presents our approach, Chapter 5 describes our experimental methodology and results, and Chapter 6 and 7 present future work and conclusions.

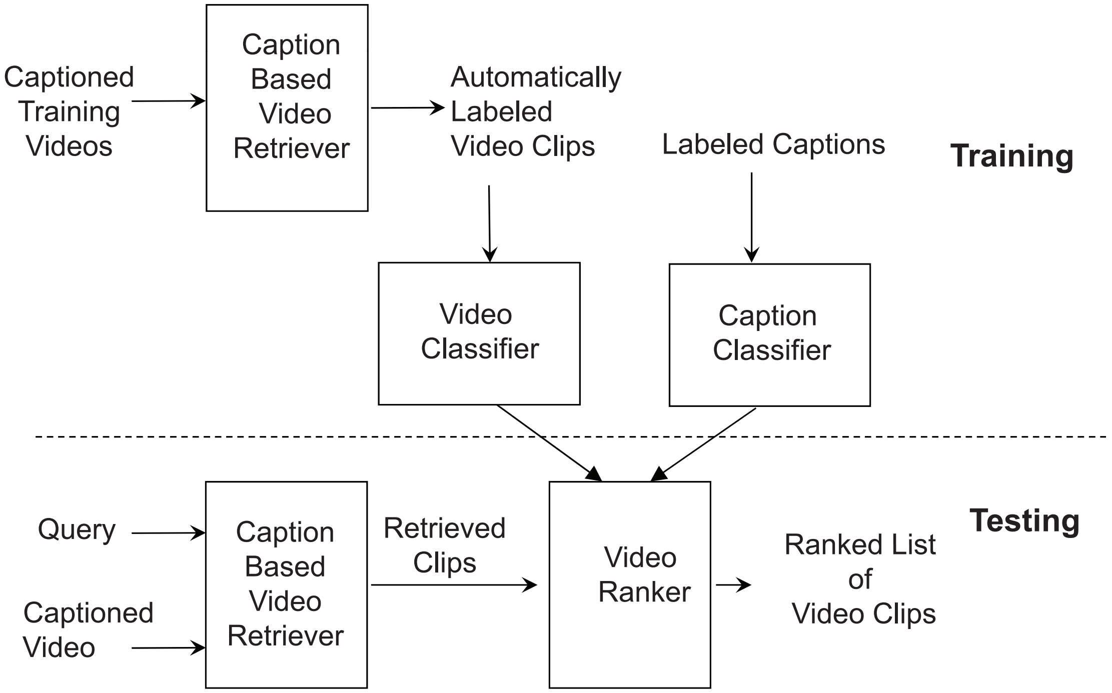

Captioned Training Videos
Caption Based Video Retriever
Automatically Labeled Video Clips
Labeled Captions
Training
Video Classifier
Caption Classifier
Query
Captioned Video
Caption Based Video Retriever
Retrieved Clips
Video Ranker
Ranked List of Video Clips
Testing

Figure 1.2: An overview of our video retrieval system

Chapter 2

Related Work

Activity recognition in videos has attracted significant attention in recent years. In the past, activity classifiers were trained mainly using human labeled video clips. Recently, there has been increasing interest in using textual and audio information along with visual information for various tasks. Textual information can be acquired from closed captions, scripts or meta-data such as tags, associated text on the webpages. The textual cues can be used for improving video classification, indexing and retrieval. In addition, they provide useful indication of labels of the video clips. Researchers have recently begun to use such textual cues for obtaining labels of associated videos. Closed captions is an interesting source given its ready availability, but is nonetheless challenging due to the loose association between the caption and video, and the inherent ambiguity of text. The activity recognition systems in literature can be mainly subdivided on the basis of supervision needed for acquiring labels of videos. Supervision required by a activity recognition system is a crucial issue because human labeling of videos is a very labor intensive task, and as there has been a steep increase in amount of video generation, we need systems to automatically annotate them. Next, we describe and contrast the related work according to the supervision required by the systems.

Human Supervision

Historically, human supervision is used for obtaining labels of video examples. Many researchers have developed activity recognizers using only visual cues and hand-labeled video clips while training a video classifier (Blank et al., 2005; Efros et al., 2003; Ke, Sukthankar, & Hebert, 2007; Schuldt et al., 2004; Wang, Sabzmeydani, & Mori, 2007). Most of the activity classifiers can be broadly categorized into local and global approaches, described in detail in Section 3.1. Textual information can also be incorporated by either training a multi-modal classifier that uses both text and visual cues, or training a text classifier that classifies a clip just on the basis of textual cues, again obtained from human labeled video clips. Labels can also be acquired using semi-supervised classification techniques that require a small set of labeled data and a large set of unlabeled data. The labeled data is mostly obtained from human supervision. Gupta, Kim, Grauman, and Mooney (2008) used captions and visual information in sports video as two views for semi-supervised classification with co-training. Co-training assumes that each example is described using two different feature sets that provide different, complementary information about the instance. Closed captions and visual information can act as two different 'views' for co-training. Wang et al. (2007) proposed a semi-supervised recognition model using latent topic models, where each frame in a video sequence corresponds to a 'word'. Though, human labeling is very reliable, it is costly and time consuming.

Closed Captions and Additional Cues

Closed captions and additional cues, such as scripts and audio information, can provide ‘weak’ labels for the video clips. Though the labels obtained are noisy, they provide enough information to build a useful classification or retrieval system. Everingham, Sivic, and Zisserman (2006), Laptev et al. (2008) and Cour et al. (2008) 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. An example of script text is shown in Table 1.1. Scripts provide detail description of scenes and actions. These methods thus cannot be used for domains such as sports videos that do not have associated scripts. Laptev et al. (2008) used captions and scripts 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. Marszalek, Laptev, and Schmid (2009) exploited contextual relationships between activities and static objects like car, trees to improve accuracy of activity recognition and object detection. Cour et al. (2008) parsed a video into a hierarchy of shots and scenes using the video’s script and closed captions. They then built a generative model for scene segmentation, alignment and shot threading. Their work is again focussed on videos that have associated scripts. Wang, Duan, Xu, Lu, and Jin (2007) use co-training to combine visual and textual ‘concepts’ to categorize TV ads. They retrieved text from videos using OCR and used external sources to expand the textual features.

Closed Captions

Closed captions alone can also provide weak supervision for obtaining labels for video clips. It is especially important for videos that do not have associated scripts but have easily available closed captions. Recent work by Fleischman and Roy is the most closely related prior research. Fleischman and Roy (2007a) used both captions and motion descriptions for baseball video to retrieve relevant clips given a textual query. Additionally, Fleischman and Roy (2007b) 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 for preprocessing videos required collecting human-labeled video clips. Babaguchi et al. (2002) suggested event-based video indexing using collaborative processing of visual and closed caption streams of sports videos. Their approach requires domain knowledge of the sport to construct a tree structure required for describing events and the sequence of keywords related to an event. Nitta, Babaguchi, and Kitahashi (2000) annotated sports video by associating text segments with image segments. Their approach uses prior knowledge of the game and the key phrases generally used in its commentary. Many researchers have worked on associating objects and scenes in closed captioned news videos. Ozkan and Duygulu (2006) associated news videos with words to perform scene and object recognition, but used keyframes for recognition and thus did not use motion cues. Duygulu and Hauptmann (2004) associated news videos with words and improve video retrieval performance using clustering of shots and co-occurrence metric. Their work is expected to improve correspondence accuracy between videos and captions. This approach is difficult to work with videos recorded in less structured setting because clustering of shots in videos with sheer variety in scale, zoom, background noise, such as sports videos, can result in highly inaccurate clusters. Also, they used color histogram as a cue for clustering, which cannot be used for activity recognition in sports videos since nearly all sports-related shots have similar color histograms. Another interesting application of using closed captions in TV broadcasts proposed by Buehler, Everingham, and Zisserman (2009) is to learn sign language in TV videos using weakly aligned closed captions.

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 explicit event descriptions, but 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.

Chapter 3

Background

3.1 Activity Recognition in Videos

In this section, we will introduce two main types of approaches in activity recognition, and describe in detail the recognition system we use in our work. Action or activity recognition in videos has similar problems as object recognition in static images, such as illumination, different views, appearance and occlusion. Apart from that, camera motion, zoom and quick change in the viewpoint add difficulty to the problem. However, motion in a video can also act as an additional cue. For example, the difference between jogging and running could be captured by taking variations in the time axis into account. Most of the approaches proposed in the literature for activity recognition can be broadly divided into local patch based and holistic approaches. Holistic approaches rely on global information like silhouettes, body shapes, three dimensional shapes (e.g. Gorelick, Blank, Shechtman, Irani, & Basri, 2007; Wang & Suter, 2007; Bobick & Davis, 2001). These approaches require building complex models for recognizing body shapes and building three dimensional models. On the other hand, local based approaches use information from local patches and model significant variation in those patches (e.g. Schuldt et al., 2004; Laptev, 2005; Willems, Tuytelaars, & Gool, 2008). These models can provide a compact yet effective solution to action recognition. Figure 3.1 shows

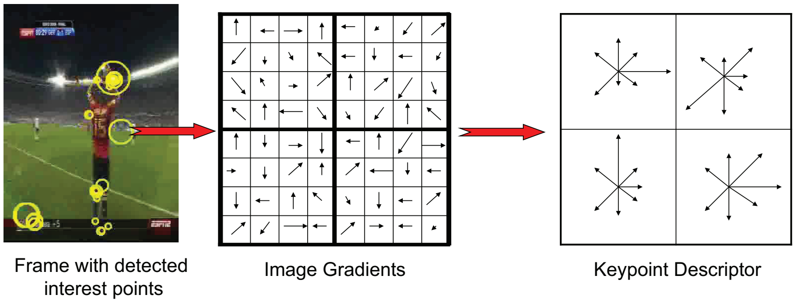


Figure 3.1: An example of local based interest point detector and descriptor.

an example of interest point detection and description using a keypoint descriptor, such as SIFT (Lowe, 2004). SIFT descriptor is scale and rotation invariant, and partially invariant to illumination changes, camera viewpoint, occlusion and clutter. Selecting one approach over another is dependent on the dataset as holistic and local approaches emphasize different aspects of activities (Sun, Chen, & Hauptmann, 2009). Local approaches are known to work better with datasets with high background noise and clutter as they focus on local motion instead of the figure shape. Holistic approaches generally work better with datasets with less background noise and more inter-class similarity as they focus on global information like figure shape.

Laptev (2005) introduced a local descriptor based approach for activity recognition and later extended it in (Laptev et al., 2008). To detect spatio-temporal events, Laptev et al. (2008) builds on Harris and Forstner's interest point operators (Forstner & Gulch, 1987; Harris & Stephens, 1988) and detects local structures where the image values have significant local variation in both space and time. They estimate the spatio-temporal extent of the detected events by maximizing a normalized spatio-temporal Laplacian operator over multiple spatial and temporal scales. Specifically, the extended spatio-temporal "cornerness" H at a given point is computed as introduced in Laptev (2005):

$$\mu = g(.;\sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} \tag{3.1}$$

$$H = \det(\mu) - k\ trace^3(\mu), \tag{3.2}$$

where '$*$' represents convolution, $g(.;\sigma_i^2, \tau_i^2)$ is a 3D Gaussian smoothing kernel with a spatial scale σ and a temporal scale τ, and L_x, L_y and L_t are the gradient functions along the x, y and t directions, respectively. In Equation 3.1, μ represents a second order spatio-temporal matrix. The points that have a large value of H are selected as interest points. The interest points can be *described* by either Histogram of Oriented Gradients, Histogram of Optical Flow, or both. We use this approach to activity recognition for describing the activities in our dataset. We choose 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. We use bag-of-words approach for representing each video clip, as described in Section 4.2.

3.2 Ensemble Learning

We use DECORATE, a ensemble classifier, for classifying video clips in our system. Ensemble Learning combines multiple learned models under the assumption that a diverse committee of learned models produces more accurate results.

The final prediction on a test example aggregates the predictions of these multiple learned models. Boosting (Freund & Schapire, 1996) and bagging (Breiman, 1996) are two popular approaches. Bagging uses different random subsamples of the training data to generate multiple classifiers. Boosting, on the other hand, iteratively learns a sequence of classifiers, each one trained to overcome the mistakes of the previous one. At each iteration, it assigns higher weights to examples that are misclassified by the previous classifier, and adds a new classifier to the committee trained on the reweighted examples.

DECORATE (Melville & Mooney, 2003) is an ensemble classifier that has been shown to be superior to boosting and bagging when learning from training sets that are small and/or noisy (Melville, Shah, Mihalkova, & Mooney, 2004). It generates a diverse ensemble of classifiers using additional artificially-constructed training examples. The algorithm generates ensembles iteratively using both the original training data and a set of artificial data labeled by the algorithm itself in the previous step. At each iteration, artificial data is generated from a simple probabilistic model generated from the training data. These artificial examples are labeled such that the true labels differ maximally from the predictions of the current ensemble. This ensures that the next member of the ensemble will disagree with the current ensemble on these examples. The result is a very diverse committee that generalizes well and prevents over-fitting. DECORATE is appropriate for our problem setting because the labeled data are automatically obtained from captioned videos and the labels are very noisy.

3.3 Subsequence Kernel

We use a subsequence kernel based classifier for training the caption classifier in our system. Kernel-based methods such as *support vector machines* (SVMs), allow using a linear classifier to solve a non-linear problem by implicitly mapping the examples into a higher dimensional space where they become linearly separable (Cristianini & Shawe-Taylor, 2000). A *kernel* computes an inner product between the mapped data points in a higher dimensional space. Intuitively, a kernel defines a measure of similarity between two examples in this higher-dimensional space. A subsequence kernel (Lodhi et al., 2002) measures the similarity of two strings by computing an inner product in the feature space composed of all possible subsequences. A subsequence is any ordered sequence of tokens occurring either contiguously or non-contiguously in a string. The similarity of two sequences is defined as the total number of subsequences that they share. For example, the phrase ‘time flies like an arrow’ is more similar to ‘time moves quickly just like an arrow’ than ‘an arrow moves quickly just like time’, and this similarity can be captured by a subsequence kernel, which will find the common subsequence ‘time like an arrow’ between the first two phrases. The subsequences are weighted by an exponential decay factor that penalizes longer subsequences.

The subsequences can be over multiple information sources, for example words, characters, part-of-speech (POS) tags. Bunescu and Mooney (2005) proposed a generalization of subsequence kernels that integrates information from multiple subsequence patterns, in the following way. Let $\Sigma_1, \Sigma_2, ..., \Sigma_k$ be several disjoint feature spaces. In our work, Σ_1 is the set of words and Σ_2 is the set of POS tags. Let $\Sigma_\times = \Sigma_1 \times \Sigma_2 \times ... \times \Sigma_k$ be the set of all possible feature vectors, where a feature vector is associated with each word in a sentence. Given two feature vectors $x, y \in \Sigma_\times$, let $c(x, y)$ denote the number of common features between x and y. Let s, t be two sequences over the finite set $\Sigma_\times$, and let $|s|$ denote the length of $s = s_1...s_{|s|}$. The sequence $s[i:j]$ is the contiguous subsequence $s_i...s_j$ of s. Let $\mathbf{i} = (i_1, ..., i_{|\mathbf{i}|})$ be a sequence of $|\mathbf{i}|$ indices in s, in ascending order. The length $l(\mathbf{i})$ *of the index sequence* $\mathbf{i}$ *in* s is defined as $i_{|\mathbf{i}|} - i_1 + 1$. Similarly, **j** is a sequence of $|\mathbf{j}|$ indices in t.

Let $\Sigma_\cup = \Sigma_1 \cup \Sigma_2 \cup ... \cup \Sigma_k$ be the set of all possible features. The sequence $u \in \Sigma_\cup^*$ is a (sparse) subsequence of s if there is a sequence of $|u|$ indices $\mathbf{i}$ such that $u_k \in s_{i_k}$, for all $k = 1, ..., |u|$. Equivalently, $u \prec s[\mathbf{i}]$ is defined as a shorthand for the component-wise '$\in$' relationship between u and $s[\mathbf{i}]$.

They define $K_n(s, t, \lambda)$, shown in Equation 3.3, as the number of weighted sparse subsequences u of length n common to s and t (i.e. $u \prec s[\mathbf{i}]$, $u \prec t[\mathbf{j}]$), where the weight of u is $\lambda^{l(\mathbf{i})+l(\mathbf{j})}$, for some $\lambda \leq 1$. λ is a decay factor that penalizes longer subsequences.

$$K_n(s, t, \lambda) = \sum_{i:|i|=n} \sum_{j:|j|=n} \prod_{k=1}^{n} c(s_{i_k}, t_{j_k}) \lambda^{l(i)+l(j)} \tag{3.3}$$

Since subsequences of words take word order into account, a subsequence kernel can exploit syntactic cues unavailable to a standard unordered "bag of words" text classifier (Sebastiani, 2002); therefore, we found in our experiments that it obtained superior accuracy for determining caption relevance. In our example above, a bag-of-words based similarity measure will give equal score to both ‘time moves quickly just like an arrow’ and ‘an arrow moves quickly just like time’.

Chapter 4

Approach

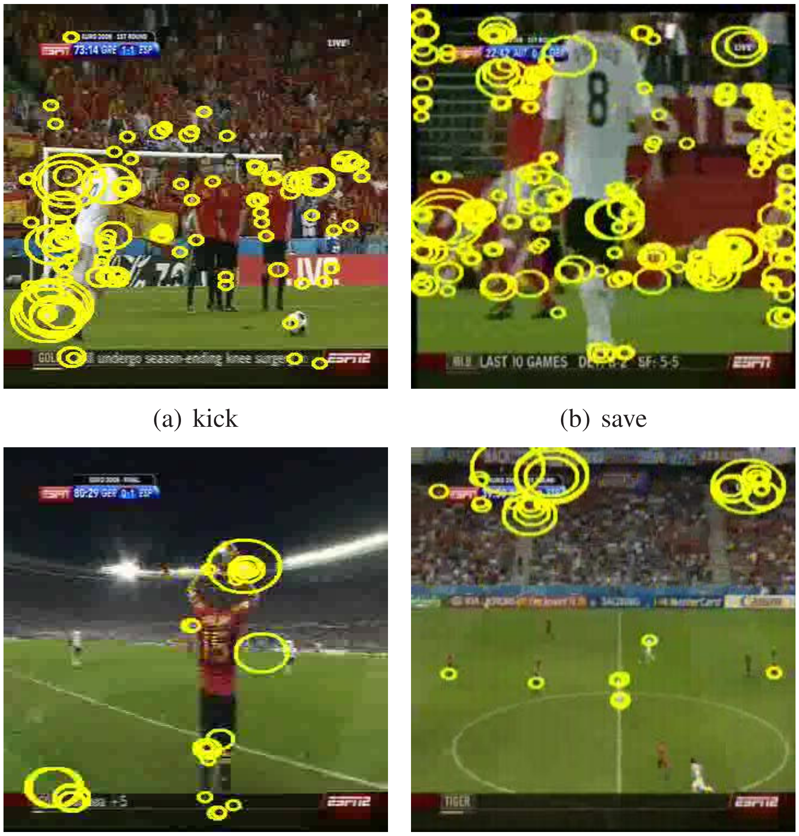
We first describe our procedure for automatically collecting labeled clips from captioned videos. We then explain the encoding of videos using motion descriptors and how to use them to train a video classifier. Next, we describe our caption classifier, and finally we explain the overall system for retrieving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

4.1 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 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%, and thus constraining the system to choose clips with exactly one keyword in their captions does not affect the system much. We then extract a fixed-length clip around the corresponding time in the video. In our dataset, we qualitatively found that extracting 8 second clips mostly captures activities in the videos. 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.

4.2 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) (see Section 3.1). These features are shown to have worked well with human activity recognition in real-world videos (Laptev & Perez, 2007; Laptev et al., 2008; Marszalek et al., 2009). In addition, this approach can be used for detecting activities in many domains as it does not



(a) kick

(b) save

(c) throw

(d) touch

Figure 4.1: Example frames from the four query classes with detected motion features

use any domain-specific features or prior knowledge of the game.

As described in Section 3.1, 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 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 4.1 shows example frames of query classes with detected motion features. We can see that the motion features are detected mostly on the interesting and useful patches. However, when the players are very small in size and there is background clutter, many interest points are detected in the background as well.

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 & Frank, 2005), 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 (Melville & Mooney, 2003; Melville et al., 2004) (see Section 3.2).

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. We build 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. We also tried multiclass classifiers, but they gave inferior performance. This is expected since generally binary classification (one-against-all) performs better than multiclass (one-against-one) classification.

Our approach creates a category model and one can retrieve and rank video clips in a dataset for each category concept. In a real life application system, the system needs to have pre-specified finite number of categories, which is realistic in most domains.

4.3 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 4.1. We expect the system to learn that subsequences like ‘last game’, ‘this weekend’, ‘needed touch’ are irrelevant to events going on in the video, and subsequences like ‘earns kick’, ‘first touch’, ‘gets ball’, ‘conceding this time’ are relevant. As explained in Section 3.3, a subsequence ker-

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
Mcgeady gets the ball and works it into a nice shot and Van der sar comes across and makes a beautiful save.	1
They scored in the last kick of the game against the Czech Republic.	0
Got kicked in the face.	0
And Dempsey , with the first touch.	1
Gary Neville conceding the throw this time.	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
Pushed ahead, Bradley .	1
Galaxy and other teams missing prominent player this weekend because of world cup qualifying.	0
Mertesacker getting in the way.	1
Conversation going on .	0
Throwing here for cuba .	1
Take your time when you are throwing.	0
Beautifully placed to Philipp Lahm.	1

Table 4.1: Some examples of captions with their labels in our dataset. Label '1' means that the caption is relevant to some event in the game.

nel is apt for learning such subsequences, which otherwise is not captured by most commonly used bag-of-words approach. An n-gram model, which uses n-1 words of prior context, might perform better than bag-of-words in this scenario. N-gram models have been used in speech recognition, OCR recognition etc. But we expect a subsequence kernel to outperform an n-gram based approach. The reason is that an n-gram approach cannot skip words and needs proper smoothing for rarer phrases. For example, the number of common 2-grams in 'kick ball' and 'kick the ball' is zero. On the other hand, the phrases share the subsequence 'kick ball', though the similarity score of the subsequence kernel will be penalized for skipping a word.

We use an SVM string classifier that uses a subsequence kernel (Lodhi et al., 2002), which measures the number of subsequences shared by two strings (see Section 3.3). We use two subsequence patterns: word subsequences and Part-of-Speech (POS) subsequences. The Stanford POS tagger (Toutanova, Klein, Manning, & Singer, 2003) was used to obtain POS tags for each word and we used LibSVM (Chang & 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. 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. One might expect to need both video and text association while labeling the captions but as can be seen from the captions in Table 4.1, labels of the captions are pretty intuitive, especially when labeling them sequentially.

4.4 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 4.1. 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 ($P(label|clip)$) 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):

$$\begin{aligned} P(label|clip\ with\ caption) &= \alpha P(label|clip) \\ &+ (1-\alpha)P(relevant|clip\text{-}caption) \end{aligned} \tag{4.1}$$

The value of α is determined empirically as described in Section 5.2.

Chapter 5

Experiments

5.1 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 1.1. There is a wide difference in camera angle and zoom among the clips for a category. Sometimes, the players are so small that even humans have difficulty in labeling the clips. Also, in some clips, the activity is occluded and the background noise is very high. Compare the examples of our dataset in Figure 1.1 from the ones from KTH and Weizmann dataset shown in Figures 5.2 and 5.3. KTH and Weizmann datasets are recorded in simplified settings with little or no camera motion and the size of objects do not vary much. We extracted clips for four activity keywords: $\{kick, save, throw, touch\}$, as discussed in Section 4. The total number of clips extracted was 624. 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. For each category class, a set of video clips are retrieved from the video dataset, and they are labeled with the category label if they depict the corresponding activity, irrespective of the fact that some of them have multiple activities. For example, if a clip is retrieved for category class ‘save’ and it has both ‘save’ and ‘kick’ activities, it is labeled ‘save’. However, if the clip depicts only a kicking event, it is labeled ‘kick’. If a clip doesn’t depict activities any of the categories, then it is labeled *incorrect*. The system itself never uses these gold-standard labels. Figure 5.1 shows the the number of correct and incorrect clips for each class. Note that the automatically labeled data extracted using captions is extremely noisy. We can see that the noise level (percentage of clips that are not correct) is particularly high for classes kick and throw. The query class ‘kick’ has most noise, interestingly because apart from unrelevant captions, ‘kick’ word is generally used to convey two meanings in soccer commentary: kicking of a ball and kicking of a person. We are considering the former meaning for the query ‘kick’.

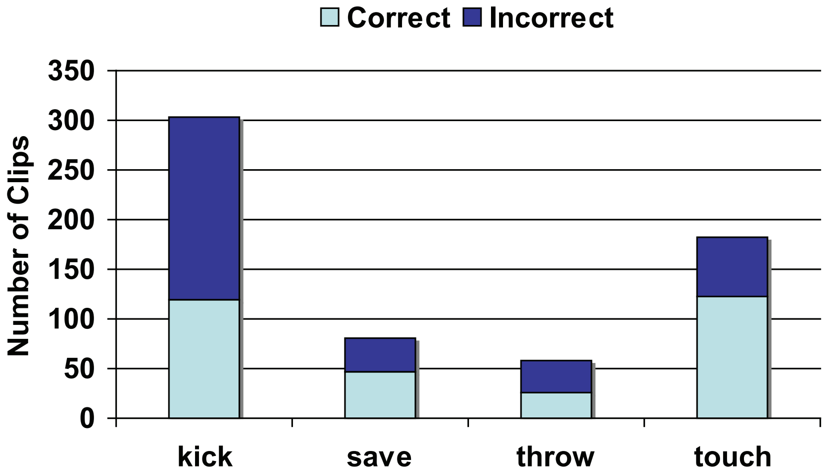


Figure 5.1: The number of total number of clips for each category, and indicating the number of correct and incorrect clips

The caption classifier is trained using a disjoint set of four games. Each

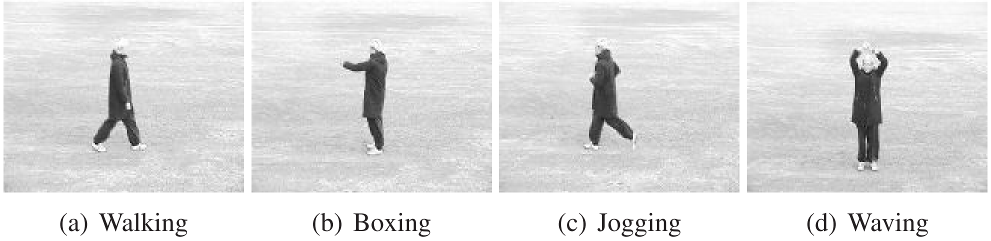

(a) Walking (b) Boxing (c) Jogging (d) Waving

Figure 5.2: Examples from KTH dataset for four classes. Note that all examples are recorded in simplified and less noisy settings.

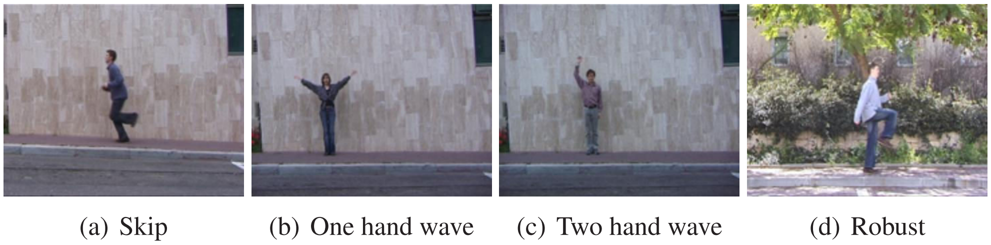

(a) Skip (b) One hand wave (c) Two hand wave (d) Robust

Figure 5.3: Examples from Weizmann dataset. There is no camera motion and there is very less viewpoint and zoom change.

sentence in the text commentary of these games is manually labeled as *relevant* or *irrelevant* to the current activity in the game. To reduce human time and effort, this labeling is performed without examining the corresponding video. All 4,368 labeled captions in this data are used to train the caption classifier. The dataset consists of 1,371 captions labeled as *relevant*.

5.2 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 4.1, 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 also tried selecting different α for different classes but unexpectedly it gave worse performance. The intuition behind learning different α for each class is that for some classes the video classifier might perform better than the caption classifier and vice-versa for the other classes.

For each query ($kick$, $save$, $throw$, $touch$), we retrieve and rank clips in the test game as explained in Section 4.4. 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 (Manning, Raghavan, & Schütze, 2008). If the set of retrieved clips for a query $q_i \in Q$ is $\{clip_1, clip_2, ..., clip_{m_i}\}$ 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). BASELINE doesn't depend on the video classifier. 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 positive examples as determined by the human labeling (called GOLD-VIDEO).

5.3 Results

Ranking using the Video Classifier

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 5.1: Retrieval Results: MAP scores when ranking the retrieved clips using a video classifier.

Classifier	DECORATE	SVM
Majority Class Baseline	33.9	33.9
VIDEO	19.8	28.18
GOLD-VIDEO	20.4	31.30

Table 5.2: Classification Results: Macro-average F-measure of the video classifiers and baseline when classifying video clips.

Table 5.1 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. 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 rest of the retrieval results, the video classifier is trained with DECORATE since it performs the best.

Video Classifier The video classifier can also be used for classifying video clips. Table 5.2 shows macro-average F-measure (see Manning et al., 2008) of the classifiers in one-against-all class classification with leave-one-game-out cross validation. The majority class baseline means all clips are labeled with the class that has the most number of examples, and it is stronger than the random baseline. For example, if in a binary classification task there are 70% negative examples, then the majority class baseline will give 70% accuracy, as it will label all clips as negative. As can be seen, the video classifier performs worse than the majority class baseline; however as shown earlier, it is still useful for improving ranking of clips within each class. For classification, SVM performs better than DECORATE and Gold-Video performs better than Video. This shows that the probability predictions by DECORATE used for ranking are better than SVM even if the binary classification predictions are worse.

Ranking using the Caption Classifier

As explained in Section 4.4, the caption classifier can also be used to rank results. The MAP score for ranking with the caption classifier is shown in Table 5.3.

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

Table 5.3: MAP measures for different approaches

Approach	Accuracy
Majority Class Baseline	69.02
Bag-of-Words	69.07
WORD SSK	79.26
WORD+POS SSK	79.81

Table 5.4: Classification accuracy of the caption classifier, when trained and tested using leave-one-game-out on labeled captions from four games (which are disjoint from the primary dataset)

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. It is interesting to note that VIDEO and CAPTION perform almost the same, although they are trying to capture different aspects.

Caption Classifier 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. As Table 5.4 shows, the classification accuracy of an SVM with a subsequence kernel that includes word and POS subsequences. (WORD+POS SSK) is 79.81%, compared to a subsequence kernel that uses just word subsequences (WORD SSK) and to a baseline of 69.02%

Sentence	Label
That bull is hit deep into left center field.	1
Penny used to be much more of a strikeout pitcher.	0
A fastball for a strike.	1
When he went out of the lineup, he was batting.	0
Jason Marquis , only the first batter he faced has been able to reach base against him.	0
Penny throws the fastball on the inside corner and Derosa hits the line drive right out of here.	1
I think he is under rated as a pure hitter.	0
And not a bad pitch.	1

Table 5.5: Some examples of captions from the baseball test set. Label '1' means that the caption is relevant to some event in the game.

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 5.6: Classification accuracy of the caption classifier, when trained on soccer captions and tested on baseball captions

when all captions are labeled with the majority class. The results also show that the subsequence kernel that uses just the words outperforms the baseline by around 10% accuracy. Using an SVM with a bag-of-words approach gave similar results as the baseline, signifying the importance of word order and subsequences. In our video retrieval experiments, we used WORD+POS subsequence kernel, though we expect similar results with WORD subsequence kernel.

To show generality of the caption classifier across different datasets, we use the caption classifier to classify baseball captions when it was trained on soccer captions and optionally a small number of baseball captions. The content of both sports commentary are quite different as seen from Table 4.1 and Table 5.5, as sportscasters tend to use sports-specific words and comments. The baseball dataset consists of 985 hand labeled captions from a baseball game labeled as 'relevant' or 'irrelevant'. Table 5.6 shows results of *transfer learning* from soccer captions to baseball captions. Transfer learning is generally performed using a large set of out-of-domain data and a small or empty set of in-domain data. The training set consisted of all the labeled soccer captions (4368 examples) and a part of baseball captions dataset. The baseball dataset was split and tested using five-fold cross validation (that is, in each iteration, data in four folds was added to the training dataset and the other fold was used for testing). In the table, 'Soccer and Baseball' refers the case when both soccer and baseball captions are used while training the classifier. 'Soccer' refers when only soccer captions (out-of-domain dataset) was used while training and similarly 'Baseball' refers when only baseball captions (in-domain dataset) was used. WORD SSK using Soccer and Baseball dataset performs the best signifying the importance of word order and transfer learning. On the other hand, using subsequences of Part-of-Speech tags hurts the performance as WORD+POS SSK doesn't perform well. As can be seen, WORD SSK using the Baseball dataset performs worse than the baseline because it had very few examples in the training set. It is interesting to note that WORD SSK using only the Soccer dataset performs better than most of the other classifiers even though the training set does not contain a single example from the baseball domain. Table 5.7 shows a part of test results of classifying baseball captions along with the captions true labels, when the classifier is trained only on soccer captions. We can see that the subsequence kernel predicts captions with subsequences such as ‘left center’, ‘runner going’, ‘hits line’, ‘watch upper body’ as *relevant* and with subsequences such as ‘sunday night’, ‘has been able’, ‘this year’ as *irrelevant*. The results and examples presented in Tables 5.6 and 5.7 show that the caption classifier is trying to detect a very abstract linguistic property (depiction of a current event) and it should generalize fairly well to other domains as well. An interesting mistake by the caption classifier in Table 5.7 is classifying ‘*He got hit with a shot hit right back at him by Derrek Lee.*’ as *relevant*. The caption seems relevant to humans too, however, it is *irrelevant* because the sportscaster was discussing an injury that happened in a previous game. We hope that an approach that uses preceding and succeeding captions while classifying a caption might help.

Aggregating the rankings

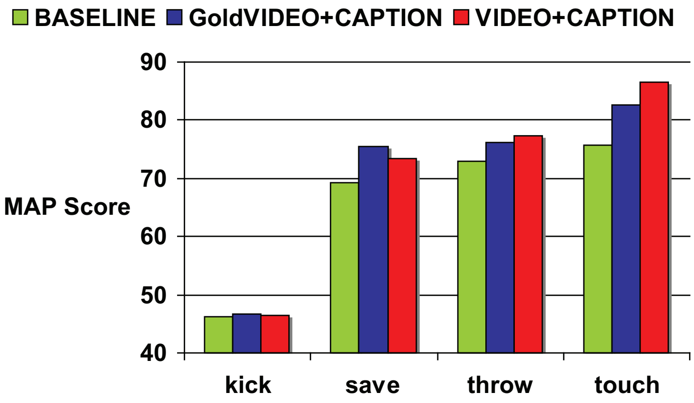


Figure 5.4: MAP scores for each query using different approaches

Sentence	True Label	Predicted Label
And to left center.	1	1
Brad Penny hit on the left toot with aline drive in the first inning.	1	1
The runner going, the ball is popped up.	1	1
Aramis Ramirez takes a ball.	1	1
And that is down the left field line.	1	1
One ball, one strike, the count to him.	1	1
Penny throws the fastball on the inside corner and Derosa hits the line drive right out of here.	1	1
Cedeno hitting a shot.	1	1
Watch the upper body.	1	1
This was during the first inning.	0	0
Kent who at age 40 can get around on anybody is fastball.	0	0
He always has been able to hit.	0	0
He has had 22 this year and only thrown out 6 times.	0	0
Sunday night baseball from dodger stadium, Los Angeles.	0	0
Throw strikes .	1	0
And Ronny Cedeno throws him out .	1	0
Marquis is a good hitter, as well .	0	1
He got hit with a shot hit right back at him by Derrek Lee.	0	1

Table 5.7: Baseball captions along with their true and predicted labels when the classifier is trained only on soccer captions.

Ranking using VIDEO MAP score = 73.33	Ranking using CAPTION MAP score = 67.91	Ranking using VIDEO +CAPTION MAP score = 80.41 $\alpha = 0.3$
✔Clip1	✖Clip7: Lovely touch.	✔Clip1: Just trying to touch it on.
↓	↓	↓
✖Clip2	✔Clip1: Just trying to touch it on.	✖Clip7: Lovely touch.
↓	↓	↓
✖Clip4	✔Clip6: Just touched on by Nani.	✔Clip6: Just touched on by Nani.
↓	↓	↓
✔Clip6	✖Clip2: If he had not touched it.	✖Clip2: If he had not touched it.
↓	↓	↓
✖Clip7	✖Clip4: I do not think it was touched.	✖Clip4: I do not think it was touched.

Table 5.8: 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.

The rankings of the video and caption classifiers leverage two different sources of information, visual and linguistic, respectively. Table 5.3 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 5.3 are statistically significant as compared to BASELINE on a one-tailed paired t-test with a 95% confidence level. The average value of α, computed by cross-validation on the held-out set (see Section 4.4 and Section 5.2), over all the games is 0.46. Figure 5.4 shows MAP scores for each of the four queries when using different approaches. Sometimes there are no correct instances of a query in a game and the corresponding MAP score becomes NaN. Note that since we ignore NaN values when averaging MAP scores across the folds of the leave-one-game-out cross-validation, the final MAP score is not exactly equal to the average of the MAP scores of the individual queries shown in Figure 5.4. We can see that VIDEO+CAPTION improves the MAP score most for the query ‘touch’, and least for ‘kick’. This is expected since noise in the automatically labeled data was highest for ‘kick’ and lowest for ‘touch’(see Figure 5.1).

Table 5.8 shows MAP scores and rankings (from most to least relevant) produced by VIDEO, CAPTION, and VIDEO+CAPTION for the query ‘touch’ for a particular test game. There were seven clips extracted from the game for the given query. Two clips got same rankings by all three approaches and are thus not shown in the table. For the test game, α computed as 0.3 was used for aggregating the rankings. As expected, the MAP score for VIDEO+CAPTION is higher than

VIDEO and CAPTION individually. The example clearly shows that the VIDEO and CAPTION classifiers leverage different information, and that aggregating them produces better results. For example, even though VIDEO incorrectly ranks Clip2 and Clip4 fairly high, CAPTION gives them low rank, thus decreasing their rank in VIDEO+CAPTION. Similarly, Clip7 was incorrectly ranked the highest by CAPTION, but VIDEO gives it a low rank, pushing its rank down when they are aggregated. Clip7, corresponding to the caption ‘lovely touch’, is not actually relevant to the query ‘touch,’ since commentators were discussing an event that happened several seconds back and the video clip did not actually capture the event.

Chapter 6

Future Work

Exploiting the multi-modal character 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 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. To further improve precision, it will be interesting to learn temporal patterns of keywords (Babaguchi et al., 2002) associated with an event from the captioned video data. 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 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.

To improve recall, we could also use the video classifier to label other video clips that do not have the query class keywords in their captions, and then use captions of the newly classified video clips to learn a text classifier that can classify captions of video clips that whether or not they are instances of the target. It will be similar to the text classifier introduced in (Laptev et al., 2008), except that this would not require any labeled video clips. This, however, requires a video classifier that would classify human activities having high clutter and background noise with high accuracy and is thus left to future research in human activity classification.

The caption classifier currently classifies each caption separately. We expect that a classifier that takes preceding and succeeding captions into account will perform better for some cases. When commentating on sports videos, sportscasters generally discuss relevant and irrelevant events in sequence and including the context might improve the classifier.

Another promising direction is to exploit temporal relations between activities to improve the video 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 clip preceding it in time is a ‘kick’. Hidden Markov Models and Conditional Random Fields are known for modeling sequences and might be used to model these temporal relations.

Chapter 7

Conclusion

In this thesis, 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. Our experiments show that DECORATE performs really well for video datasets having 'noisy' labels. In addition, we have shown that training a caption classifier to identify captions that describe current activities can improve precision even further. We also show that the caption classifier generalizes well across other sports domains. The encouraging results from aggregating video retrieval rankings from the video and caption classifiers further indicates that exploiting the multimodal nature of closed-captioned video can improve the effectiveness of activity recognition and video retrieval approaches.

Bibliography

Babaguchi, N., Kawai, Y., & 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., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005)*, Beijing, China.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2).

Buehler, P., Everingham, M., & Zisserman, A. (2009). Learning sign language by watching tv (using weakly aligned subtitles). In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, USA.

Bunescu, R. C., & Mooney, R. J. (2005). Subsequence kernels for relation extraction. In Weiss, Y., Schölkopf, B., & Platt, J. (Eds.), *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, Vancouver, BC.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

Cour, T., Jordan, C., Miltsakaki, E., & Taskar, B. (2008). Movie/script: Alignment and parsing of video and text transcription. In *Proceedings of the Tenth European Conference on Computer Vision (ECCV 2008)*, Marseille, France.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Duygulu, P., & Hauptmann, A. G. (2004). What's news, what's not? associating news videos with words. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)*, Nice, France.

Everingham, M., Sivic, J., & Zisserman, A. (2006). Hello! My name is... Buffy – Automatic naming of characters in TV video. In *Proceedings of the Seventeenth British Machine Vision Conference (BMVC 2006)*, Edinburgh, United Kingdom.

Fleischman, M., & Roy, D. (2007a). Situated models of meaning for sports video retrieval. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational*

Linguistics (NAACL-HLT-07); Companion Volume, Short Papers, Rochester, USA.

Fleischman, M., & Roy, D. (2007b). Unsupervised content-based indexing for sports video retrieval. In *Ninth ACM Workshop on Multimedia Information Retrieval (MIR 2007)*, Augsburg, Germany.

Fleischman, M., & Roy, D. (2008). Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, USA. Association for Computational Linguistics.

Forstner, W., & Gulch, E. (1987). 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*.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, Bari, Italy.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Gupta, S., Kim, J., Grauman, K., & Mooney, R. J. (2008). 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.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the Fourth Alvey Vision Conference (AVC 1988)*, Manchester, United Kingdom.

Ke, Y., Sukthankar, R., & Hebert, M. (2007). Event detection in crowded videos. In *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV 2007)*, Rio de Janeiro, Brazil.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, USA.

Laptev, I., & Perez, P. (2007). Retrieving actions in movies. In *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV 2007)*, Rio de Janeiro, Brazil.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision (IJCV)*.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research (JMLR)*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, USA.

Melville, P., & Mooney, R. J. (2003). Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003)*, Acapulco, Mexico.

Melville, P., Shah, N., Mihalkova, L., & Mooney, R. J. (2004). Experiments on ensembles with missing and noisy data. In *Proceedings of the Fifth International Workshop on Multi Classifier Systems (MCS 2004)*, Cagliari, Italy.

Nitta, N., Babaguchi, N., & Kitahashi, T. (2000). Extracting actors, actions and events from sports video - a fundamental approach to story tracking. In *Proceedings of the International Conference on Pattern Recognition (ICPR 2000)*, Barcelona, Spain.

Ozkan, D., & Duygulu, P. (2006). Finding people frequently appearing in news. In *Proceedings of the Fifth International Conference on Image and Video Retrieval (CIVR 2006)*, Tempe, USA.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, Washington, DC, USA.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Sun, X., Chen, M., & Hauptmann, A. (2009). Action recognition via local descriptors and holistic features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, USA.

Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to Learn*. Kluwer Academic Publishers, Boston.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of Human Language Technology Conference / North American Association for Computational Linguistics Annual Meeting (HLT-NAACL-2003)*, Edmonton, Canada.

Wang, J., Duan, L., Xu, L., Lu, H., & Jin, J. S. (2007). Tv ad video categorization with probabilistic latent concept learning. In *Multimedia Information Retrieval (MIR)*.

Wang, L., & Suter, D. (2007). Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Transactions on Image Processing*.

Wang, Y., Sabzmeydani, P., & Mori, G. (2007). Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. In *Second Workshop on Human Motion Understanding, Modeling, Capture and Animation*, Rio de Janeiro, Brazil.

Willems, G., Tuytelaars, T., & Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the Tenth European Conference on Computer Vision (ECCV 2008)*, Marseille, France.

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

Vita

Sonal Gupta was born in Bhopal, India on September 4 1985, the second daughter of Sudha Gupta and Arvind Kumar Gupta. She completed her twelve years of school education at Jawahar Lal Nehru School in Bhopal. After two years of hard work, she joined Indian Institute of Technology (IIT) Roorkee. She received the Bachelor of Technology degree in Computer Science & Engineering from IIT Roorkee in May 2007. She spent the summer of 2005 at Google R&D center at Bangalore. She joined the University of Texas at Austin in Fall 2007 and finished her Master's with thesis in August 2009. She spent the wonderful summer of 2008 in Redmond while interning with Microsoft Research. She is now moving to the Stanford University, CA to receive her PhD in Computer Science.

Permanent address: 43 L B Sonagiri
Bhopal, M.P., India 462021

This thesis was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.