In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), pp. 457--472, Antwerp Belgium, September 2008.

Watch, Listen & Learn: Co-training on Captioned Images and Videos

Sonal Gupta, Joohyun Kim, Kristen Grauman and Raymond Mooney

Department of Computer Sciences
The University of Texas at Austin
1 University Station C0500
Austin, Texas 78712-0233
U.S.A.
{ sonaluta,scimitar,grauman,mooney }@cs.utexas.edu

Abstract. Recognizing visual scenes and activities is challenging: often visual cues alone are ambiguous, and it is expensive to obtain manually labeled examples from which to learn. To cope with these constraints, we propose to leverage the text that often accompanies visual data to learn robust models of scenes and actions from partially labeled collections. Our approach uses co-training, a semi-supervised learning method that accommodates multi-modal views of data. To classify images, our method learns from captioned images of natural scenes; and to recognize human actions, it learns from videos of athletic events with commentary. We show that by exploiting both multi-modal representations and unlabeled data our approach learns more accurate image and video classifiers than standard baseline algorithms.

1 Introduction

Systems able to automatically annotate and index visual content will be crucial to managing the world's ever-growing stores of digital images and videos. However, learning to recognize objects and actions based on visual cues alone remains quite difficult, due to factors ranging from unpredictable illumination to the sheer variety in appearance exhibited by instances of the same class. Furthermore, accurate results often depend on access to substantial labeled data, which in practice can be cumbersome to obtain in adequate quantities.

We propose to facilitate the learning process in this domain by integrating both visual and linguistic information, as well as unlabeled multi-modal data. In particular, we consider the tasks of recognizing categories in natural scenes from images with caption text, and recognizing human actions in sports videos that are accompanied by an announcer's commentary. Both are interesting data sources given their ready availability, but are nonetheless challenging due to the loose association between the dual cues as well as the frequent ambiguity of either cue alone. We design an approach using co-training [6] that takes local appearance and spatio-temporal descriptors together with text-based features to learn the categories from a partially labeled collection of examples.

The majority of state-of-the-art systems for image and video classification use unimodal data – either visual or textual features alone [32, 12, 5, 35, 20, 14, 17]. Given the natural occurrence of both feature types together, researchers have only recently begun to explore ways to learn from multi-modal image and language data. Previous work has focused on learning the association between visual and textual information [2, 13, 11], using supervised methods to improve text-based video retrieval [15], improving audio-visual human-computer interfaces [8], and designing unsupervised methods to cluster images [3] or strengthen image features [30]. In contrast, we consider learning to classify images and videos from labeled and unlabeled multi-modal examples, demonstrating that our approach can improve the classification of novel instances by exploiting both cues—or even the visual data alone. While co-training has previously been applied to learn from two textual views [6] or two visual views [31, 7], we present comprehensive results on using visual and linguistic information as separate views, with the idea that these distinct cues will complement each another well during training.

Our main contribution is a semi-supervised approach to recognizing scenes and human actions from captioned images or commentated videos. We show that by exploiting multi-modal data and unlabeled examples, our approach improves accuracy on classification tasks relative to both unimodal and early/late fusion baselines. In addition, it yields significantly better models than alternative semi-supervised methods when only a limited amount of labeled data is available.

The remainder of the paper is organized as follows: in Section 2 we discuss related work in more detail. In Section 3 we describe our approach for extracting visual and textual features, and provide background on building a co-training classifier. In Section 5.1 we present results for learning from captioned images, while in Section 5.2 we present results for videos with commentary, and in Sections 6 and 7 we suggest future directions and present our conclusions.

2 Related Work

In previous work using captioned images, Barnard *et al.* [2] and Duygulu *et al.* [10] generat models to annotate image regions with words. Bekkerman and Jeon [3] exploit multi-modal information to cluster images with captions using an unsupervised learning technique. Quattoni *et al.* [30] describe a method for learning representations from large quantities of unlabeled images that have associated captions to improve learning in future image classification problems with no associated captions.

Many researchers have worked on activity recognition in videos using only visual cues [32, 12, 35, 20, 5]. Everingham *et al.* [13] incorporate visual information (facial and clothing matching), closed-captioned text, and movie scripts to automatically annotate faces with names in a video. They utilize textual information only for finding names of actors who are speaking at a particular time. Nitta *et al.* [28] annotate sports video by associating text segments with the image segments. Their approach is based on previous knowledge of the game and the key phrases generally used in its commentary. Fleischman and Roy [15] use text

commentary and motion description in baseball video games to retrieve relevant video clips given a textual query. Duygulu and Hauptmann [11] associate news videos with words and improve video retrieval performance. These papers focus on video retrieval rather than classification. Our results provide a novel way to incorporate text information when learning about visual human activity. Wang *et al.* [34] 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. Our paper focuses on using visual and textual features from explicitly captioned images and videos without exploiting external sources.

Co-training has previously been shown to be useful for various applications [21, 8, 31]. Levin *et al.* [23] use co-training to improve visual detectors by training two disparate classifiers. Cheng and Wang [7] suggest a new SVM algorithm called Co-SVM that uses a co-training approach and achieved better results than a normal SVM on classifying images using color and texture as separate views, and Nigam *et al.* [27] compares the effectiveness of co-training with semi-supervised EM.

However, none of the prior work has explored using low-level visual cues and text captions as two views for co-training. We present the first results showing how to learn about human activities based on both visual cues and spoken commentary, and provide a thorough evaluation of our co-training approach relative to several other relevant methods. Since image and video classification is a difficult problem and many videos and images have associated text, we believe that our co-training approach is a novel contribution to two important practical applications.

3 Approach

The main idea of our approach is to use image or video content together with its textual annotation (captions, commentary) to learn scene and action categories. To design such a system, the main components we must define are the feature representations for linguistic and static or dynamic visual cues, and the learning procedure. In this section we describe each of these elements in turn.

3.1 Visual Features

Static Image Features To describe a captioned photograph, we want to capture the overall texture and color distributions in local regions. Following [3], we compute region-based features as follows. Each image is broken into a 4-by-6 grid of uniformly sized cells. For each region, we compute texture features using Gabor filters with three scales and four orientations, and also record the mean, standard deviation, and skewness of the per-channel RGB and Lab color pixel values. The resulting 30-dimensional feature vectors for each of the 24 regions of all images are then clustered using k-Means in order to define the prototypical region responses. Each region of each image is then assigned one of k discrete

values based on the cluster centroid closest to its 30-dimensional image feature vector.

The final "bag of visual words" representing an image consists of a vector of k values, where the i'th element represents the number of regions in the image that belong to the i'th cluster. While other descriptors are certainly possible (e.g., using scale and affine invariant interest point detectors [25]), we chose these features based on their demonstrated suitability for the image-caption dataset provided in [3], which we also use in our experiments.

Motion Descriptors from Videos To represent video clips, 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 [22]. 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 builds on Harris and Forstner's interest point operators [18, 16] 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 both spatial and temporal scales. Specifically, the extended spatio-temporal "cornerness" H at a given point is computed as

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} \tag{1}$$

$$H = \det(\mu) - k\ trace^3(\mu), \tag{2}$$

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, L_z, and L_t are the gradient functions along the x, y, z, and t directions, respectively. In (1), μ represents a second order spatio-temporal matrix. The points that have a large value of H are selected as interest points.

At each interest point, we extract a HOG (Histograms of Oriented Gradients) feature [9] computed on the 3D video space-time volume. The patch is partitioned into a grid with 3x3x2 spatio-temporal blocks, and four-bin HOG descriptors are then computed for all blocks and concatenated into a 72-element descriptor. The motion descriptors from all the video clips in the training pool are then clustered to form a vocabulary. Finally, a video clip is represented as a histogram over this vocabulary, just as a static image's features are summarized by a histogram of prototypical region descriptors.

3.2 Textual Features

The text features for the images or videos consist of a standard "bag of words" representation of the captions or transcribed video commentary, respectively. We

Table 1. Co-training Algorithm

- **Inputs:** A set of labeled and unlabeled examples, each represented by two sets of features, one for each view.
- **Algorithm:** Train a classifier for each view using the labeled data with just the features for that view.
- Loop until there are no more unused unlabeled instances:
  1. Compute predictions and confidences of both classifiers for all of the unlabeled instances.
  2. For each view, choose the m unlabeled instances for which its classifier has the highest confidence. For each such instance, if the confidence value is less than the threshold for this view, then ignore the instance and stop labeling instances with this view, else label the instance and add it to the supervised training set.
  3. Retrain the classifiers for both views using the augmented labeled data.
- **Outputs:** Two classifiers whose predictions can be combined to classify new test instances. A test instance is labeled with the category predicted by the classifier with the highest confidence.

pre-processed the captions to remove stop words and stemmed the remaining words using the Porter stemmer [1]. The frequency of the resulting word stems comprised the final textual features.

3.3 Building the Classifier using Co-training

Blum and Mitchell introduced co-training, a semi-supervised learning algorithm that requires two distinct "views" of the training data [6]. It assumes that each example is described using two different feature sets that provide different, complementary information about the instance. Ideally, the two views are *conditionally independent* (i.e., the two feature sets of each instance are conditionally independent given the class) and each view is *sufficient* (i.e., the class of an instance can be accurately predicted from each view alone). Co-training first learns a separate classifier for each view using any labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data.

Co-training was initially used to classify web-pages using the text on the page as one view and the anchor text of hyperlinks on other pages that point to the page as the other view. In this work, we use the extracted visual and textual features as the two views for co-training classifiers to detect scenes and actions.

We followed the basic algorithm suggested by [6] with one additional constraint: an unlabeled example is only labeled if a pre-specified confidence threshold for that view is exceeded. The algorithm is outlined in Table 1. In each iteration, it finds the m most confidently labeled unlabeled examples for each view. If

such instances pass the threshold test, they are added to the supervised training set with the predicted label and both classifiers are retrained. The entire process continues until there are no more unlabeled instances.

4 Experimental Design

4.1 Baselines

In order to evaluate the relative strength of co-training with multi-modal data, we compare co-training with several other supervised and semi-supervised techniques that are reviewed in this section.

Early and Late Fusion Besides co-training, multi-modal fusion methods are an alternative way to utilize both sets of features. The visual and linguistic information can be 'fused' in two ways: early and late fusion [33]. In early fusion, unimodal features are extracted and then combined into a single representation. In our case, we extract visual and textual features and concatenate them into a single vector. In contrast, late fusion learns separate unimodal classifiers directly from unimodal features and then combines their results when labeling test instances. In particular, we combine the two unimodal classifiers by using the decision of the classifier with the highest confidence.

Semi-supervised EM and Transductive SVMs Semi-supervised Expectation Maximization (Semi-Sup EM) and transductive Support Vector Machines (TSVM) are two other standard approaches to semi-supervised learning. These methods can be applied to either of the two views individually, or employ both feature sets using early or late fusion.

Although typically used for unsupervised learning, Expectation Maximization (EM) can also be used in a semi-supervised setting [26]. First, Semi-Sup EM learns an initial probabilistic classifier from the labeled training data. Next, it performs EM iterations until convergence. In the E step, it uses the currently trained classifier to probabilistically label the unlabeled training examples. In the M step, it retrains the classifier on the union of the labeled data and the probabilistically labeled unsupervised examples. Semi-sup EM has typically been applied using Naive Bayes as its probabilistic classifier. For text learning, the multinomial version of Naive Bayes [24] is typically used [26]; however, for our data we found that a standard multivariate model using Gaussian distributions for continuous features gave better results. Specifically, we used the NaiveBayesSimple classifier in Weka [36].

Transductive SVMs [19] find the labeling of the test examples that results in the maximum-margin hyperplane that separates the positive and negative examples of *both* the training and the test data. This is achieved by including variables in the SVM's objective function representing the predicted labels of the unlabeled test examples. Although TSVMs were originally designed to improve performance on the *test* data by utilizing its availability during training, they

can also be directly used in a semi-supervised setting [4] where unlabeled data is available during training that comes from the same distribution as the test data but is not the actual data on which the classifier is eventually to be tested. In our experiments we evaluate the strength of our co-training approach relative to these other semi-supervised methods.

4.2 Methodology

For co-training, we use a Support Vector Machine (SVM) as the base classifier for both image and text views. We compare co-training with other supervised and semi-supervised methods, and use the Weka [36] implementation of sequential minimal optimization (SMO) [29] for SVMs (except for TSVMs as described below). SMO is set to use an RBF kernel (γ=0.01) and a logistic model to produce proper output probabilities; otherwise, default parameters are used throughout. We use a batch size of $m = 5$ for co-training. For co-training on static images, we use a confidence threshold of 0.65 for the image view and 0.98 for the text view (determined empirically through cross-validation). For video classification (where there are more classes) we use a threshold of 0.6 for the video view and 0.9 for the text view.

We evaluate all algorithms using ten iterations of ten-fold cross validation to get smoother and more reliable results. For co-training and the other semi-supervised algorithms, the test set is disjoint from both the labeled *and* unlabeled training data.

To evaluate accuracy as the amount of labeled data increases, we generate learning curves where at each point some fraction of the training data is labeled and the remainder is used as unlabeled training data. Thus, for the last point on the curve, all of the training data is labeled. With this methodology, we expect to see an advantage for semi-supervised learning early in the learning curve when there is little labeled data and significant unlabeled data. Once all of the data is labeled, we expect the predictive accuracies of semi-supervised learning and supervised learning to converge.

5 Results

This section presents our experimental results on image and video classification. Some part of our datasets and full results are available on the web at `http://www.cs.utexas.edu/users/ml/co-training`.

5.1 Learning to Categorize Captioned Images

In this section we provide results on classifying captioned static images.

(a) Caption: Ibex in Judean Desert

(b) Caption: Bedouin Leads His Donkey That Carries Load Of Straw

(c) Caption: Nuns In Seclusion Near Santa Catherina In Sinai

(d) Caption: Ibex Eating In The Nature

(e) Caption: Entrance To Mikveh Israel Agricultural School

(f) Caption: Olive Trees At The Bell Garden In Jerusalem

Fig. 1. Some images and their corresponding captions of the image dataset. Figures 1(a)-1(c) are of class 'Desert' and the rest are of class 'Trees'.

Dataset Our image data is taken from the Israel dataset[1] introduced in [3], which consists of images with short text captions. In order to evaluate the co-training approach, we used two classes from this data, Desert and Trees. These two classes were selected since they satisfy the sufficiency assumption of co-training, which requires that both views be effective at discriminating the classes (given sufficient labeled data). We refer to this set as the Desert-Trees dataset. Some examples of each class are shown in Figure 1. The complete dataset contains 362 instances. To create the vocabulary of visual words, we used k-means with k=25 (see Section 3.1). The total number of textual features for this dataset is 363.

Results and Discussion Our results comparing co-training with various other classification methods are shown in Figures 2 to 4. In the figures, "Image View" and "Text View" refers to using only the named view's features. The significance of the results were evaluated using a two-tailed paired t-test with a 95% confidence level. Based on preliminary experiments, an RBF kernel (γ=0.01) was used for the SVM in all experiments.

Comparison of Co-training to Supervised Learning. Figure 2 compares co-training using an SVM as the base classifier to supervised classification using an SVM, which is known to often be one of the best performing methods for high-

[1] `http://www.israelimages.com`

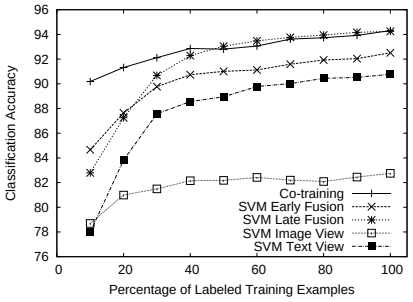
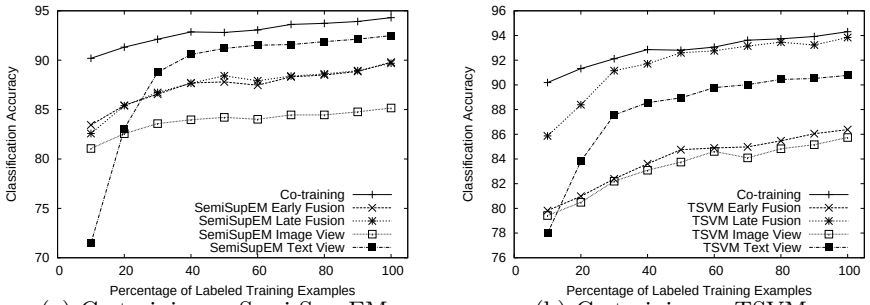


Fig. 2. Comparison of co-training with supervised classifiers on the Desert-Trees dataset. Co-training performs the best, converging with late-fusion for larger amounts of labeled data.



(a) Co-training v. Semi-Sup EM (b) Co-training v. TSVM

Fig. 3. Comparison of co-training with other semi-supervised techniques on the Desert-Trees captioned images dataset. Co-training outperforms all other methods.

dimensional data in practice. The results show that co-training is more accurate than a supervised SVM using unimodal data and early fusion of multi-modal data, with statistically significant differences at all points on the learning curve. With respect to the individual views, except at the start of the learning curve, the text view performs better than the image view. This is reasonable given that the image cues are often more indirect than the text features. The much smaller number of features in the image view allows it to do a bit better than the text view when training data is extremely limited. Both early and late fusion perform better than the unimodal classifiers since they exploit both views. Co-training is more accurate than late fusion, except for later in the learning curve where they converge. Once all the data is labeled (the last point on the learning curve), co-training and late-fusion are exactly the same since co-training has no unlabeled data to exploit.

Comparison of Co-training to other Semi-Supervised Methods. Many evaluations of semi-supervised learning only show that the proposed method performs better than supervised learning but do not compare to other semi-supervised methods [6, 19, 30]. Here we present results comparing co-training with two other

well-known semi-supervised techniques: Semi-supervised EM [26] and transductive SVMs [19]. Results are shown in Figures 3(a) and 3(b).

Figure 3(a) shows that co-training with SVM as the base classifier outperforms Semi-Sup EM irrespective of the view it considers, with statistically significant differences across the learning curve.

In order to compare with transductive SVM, we have used SVM^{light} [19], with an RBF kernel (γ=0.01) and default values for all other parameters. The figure shows that co-training performs better than transductive SVM irrespective of the view it considers. The difference in accuracy is statistically significant across the learning curve, except when compared to TSVM using late fusion. When compared to TSVM using late fusion, the difference is statistically significant when 40% or less of the training data is labeled.

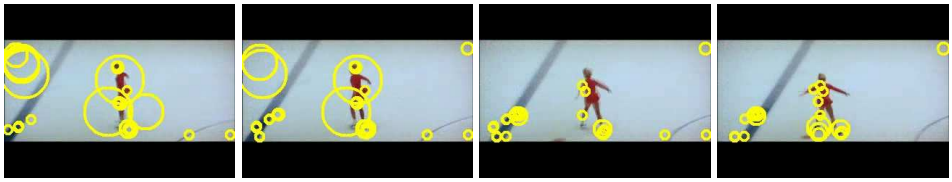
Our results are consistent with previous results on text data showing that in domains with two independent and sufficient views, co-training is more effective than Semi-Sup EM [27]. By directly exploiting the redundant nature of the visual and linguistic information, our results indicate that co-training can classify captioned images more accurately than than other semi-supervised methods.

5.2 Learning to Recognize Actions from Commentated Videos

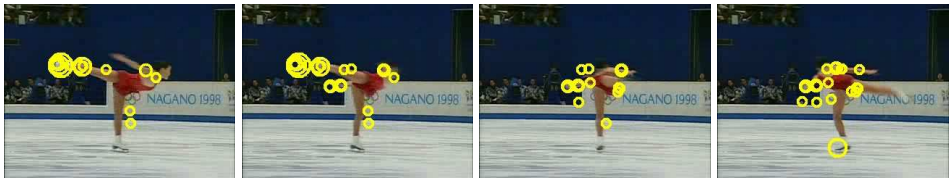
Next we report results using our co-training approach to learn human action categories from commentated videos of athletic events.

Dataset For this experiment, we collected video clips of soccer and ice skating. One set of video clips is from the DVD titled '1998 Olympic Winter Games: Figure Skating Exhibition Highlights', which contains highlights of the figure skating competition at the 1998 Nagano Olympics. Another set of video clips is on soccer playing, acquired either from the DVD titled 'Strictly Soccer Individual Skills' or downloaded from YouTube. These videos mostly concentrate on the player in the middle of the screen and usually the motions are repeated several times with different viewpoints. The soccer clips are mainly about soccer specific actions such as kicking and dribbling. There is significant variation in the size of the person across the clips.

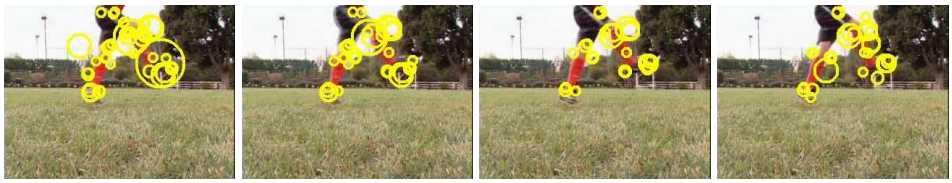
The video clips are resized to 240x360 resolution and then manually divided into short clips. The clip length varies from 20 to 120 frames, though most are between 20 and 40 frames. While segmenting activities in video is itself a difficult problem, in this work we specifically focus on classifying pre-segmented clips. The clips are labeled according to one of four categories: kicking, dribbling, spinning and dancing. The first two are soccer activities and the last two are skating activities. The number of clips in each category are, dancing: 59, spinning: 47, dribbling: 55 and kicking: 60. Example frames from each class with detected motion features and their captions are shown in Figure 4. The illustrated features are useful in discriminating between the classes and few features are detected in the background. We used k=200 in the k-means algorithm to create the vocabulary of video features (see Section 3.1).



(a) Dancing: Her last spin is going to make her win.



(b) Spinning: A female skating player is revolving in the current position many times, with her posture changing over time.



(c) Kicking: Jim uses stretches his arms outside to balance him and let goes a ferocious drive.

(d) Dribbling: The kid keeps the ball in check by juggling it with his legs.

Fig. 4. Randomly selected consecutive frames of video with detected spatio-temporal interest points. Interest points are displayed as yellow circles around the detected points. One clip per each class of dancing, spinning, kicking, and dribbling is shown above. In addition, the text commentary is also shown below each clip.

As the video clips were not originally captioned, we recruited two colleagues unaware of the goals of the project to supply the commentary for the soccer videos. The skating commentary was provided by two of the authors. Additional sample captions are shown in Figure 5. The total number of textual features is 381 for this dataset.

Results and Discussion In Figure 6 (a), we compare co-training with a supervised SVM using unimodal views and early/late fusion of multi-modal views. Co-training performs better than all other methods early in the learning curve.

Spin:
That was a very nice forward camel
Well I remember her performance last time
After gliding, she just starts to make many revolutions while maintaining her current position with her head back.
Her angular movement seems so dizzy because he spins round with her head up and down and also the movement is so fast.
Elizabeth is able to clear this one
Her beautiful performance of revolving herself makes the entire audience impressed due to her perfect posture.
Dancing:
Wow those were some great steps
He has some delicate hand movement
She gave a small jump while gliding
He does slight spins and tries to express bird's motion by dancing like it and goes forward very fast.
The crowd is cheering him a lot
She is drawing a big circle with her arms very fast while moving her body backward and shows lightweightness.
Kick:
His balance is a bit shaky but he manages to execute the kick in the end.
He runs in to chip the ball with his right foot.
He runs in to take the instep drive and executes it well.
He plants his ankle level with the ball and swings though to get the kick and makes sure he has his eyes on the ball all the time.
He come from behind and hits the rolling ball with power just as it rolls in front of him.
He runs behind the ball and has to stretch himself to kick the ball with the inside of his toes.
Dribbling:
Again the striker turns around effortlessly and kicks the ball away from the defender making it look too easy.
At fast speed as the ball is juggled between the legs it becomes difficult to control it.
The small kid pushes the ball ahead with his tiny kicks.
He does the scissors over the ball quickly to move the ball ahead.
He takes the ball with him by alternately pushing the ball forward and swinging the leg over it and using the other leg to distract the defender.
Ran uses the combination of right leg scissor and roll to take the ball ahead

Fig. 5. Captions of some video clips in the four classes

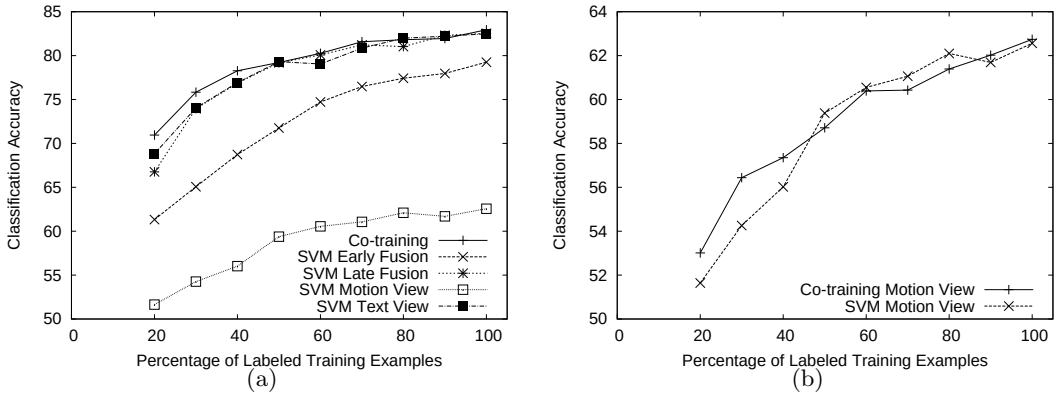


Fig. 6. (a) Comparison of co-training with early fusion, late fusion, motion view and text view on the commentated video dataset. Co-training performs better when only a small fraction of labeled data is available. (b) Co-training compared with supervised learning when text commentary is not available during testing. Co-training performs better when few labeled examples are available.

This demonstrates that utilizing unlabeled data and multi-modal views improves accuracy when supervised data is limited, a valuable advantage. Both co-training and late fusion exploit both views of the dataset, but co-training outperforms late fusion since it also uses the unlabeled data to improve accuracy. It is interesting that early fusion actually performs worse than supervised learning using the text view; we attribute this to the higher-dimensional feature vector, which increases the complexity of learning and impairs generalization.

In many real-life situations, we may not have textual commentary on the novel test videos that we wish to classify. However, even in cases where commentary is not available at test-time, we would still like to benefit from the commentary that was available during training. Therefore, we also examine the case where text is unavailable during testing and an instance must be classified using only video input. Figure 6(b) compares co-training using only the motion view during testing with a supervised SVM using the motion view. In this case, co-training performs better than SVM when only a few labeled examples are available. We also evaluated an analogous situation with the image dataset, but in that case results were comparable to a supervised SVM. All the results are statistically significant until 30% of the data is labeled.

6 Future Work

The image test corpus used in the current experiments is fairly small and requires only binary classification. We would like to test multi-modal co-training

on a larger multi-class corpus of captioned images. We would also like to extend our approach to images that do not have explicit text captions but are surrounded by related text. In particular, images on the web rarely come with explicit captions; however, it is natural to use surrounding text productively to find relevant images. By automatically extracting the appropriate surrounding text as a "pseudo-caption," multi-modal co-training could be used to improve the classification of web images. The video commentary in our experiments was added specifically for this project, although we strove to make it natural. In the future, we hope to expand our results to include video with existing closed-captioned commentary and automate the segmentation of video into clips.

7 Conclusion

Recognizing scenes in images and actions in videos are important, challenging problems. We have proposed a solution that uses co-training to exploit both visual and textual features from labeled and unlabeled data to improve classification accuracy. Our results show that such multi-modal co-training can outperform several other standard learning algorithms. By exploiting the redundant information inherent in images or videos and their textual descriptions, we have shown that the amount of supervision required to accurately classify images and videos can be significantly reduced.

8 Acknowledgment

We thank Ivan Laptev for releasing his code for computing spatio-temporal motion descriptors. This work was funded by grant IIS-0712907 from the U.S. National Science Foundation. The second author is additionally supported by a Samsung Scholarship.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.
2. Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
3. Ron Bekkerman and Jiwoon Jeon. Multi-modal clustering for multimedia collections. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-2007)*. IEEE Computer Society, 2007.
4. K. Bennett and A. Demiriz. Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, 11:368–374, 1999.
5. Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV-2005)*, pages 1395–1402, 2005.

6. Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.
7. Jian Cheng and Kongqiao Wang. Active learning for image retrieval with Co-SVM. *Pattern Recognition*, 40(1):330–334, 2007.
8. C. Mario Christoudias, Kate Saenko, Louis-Philippe Morency, and Trevor Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 84–91, New York, NY, USA, 2006. ACM.
9. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.
10. Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, London, UK, 2002. Springer-Verlag.
11. Pinar Duygulu and Alexander G. Hauptmann. What's news, what's not? associating news videos with words. In *ACM International Conference on Image and Video Retrieval (CIVR-2004)*, pages 132–140, 2004.
12. Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, Nice, France, 2003.
13. Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! My name is... Buffy – Automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006.
14. Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 12, page 178, Washington, DC, USA, 2004. IEEE Computer Society.
15. Michael Fleischman and Deb 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*, pages 37–40, Rochester, New York, April 2007. Association for Computational Linguistics.
16. Wolfgang Forstner and Eberhard 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*, pages 281–305, 1987.
17. Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.
18. Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.
19. Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 200–209, Bled, Slovenia, June 1999.
20. Yan Ke, Rahul Sukthankar, and Martial Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, October 2007.
21. Svetlana Kiritchenko and Stan Matwin. Email classification with co-training. In *Proceedings of CASCON-2001*, pages 192–201, Toronto, Canada, 2001.

22. Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.
23. Anat Levin, Paul Viola, and Yoav Freund. Unsupervised improvement of visual detectors using co-training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV-2003)*, page 626, Washington, DC, USA, 2003. IEEE Computer Society.
24. Andrew McCallum and Kamal Nigam. A comparison of event models for naive Bayes text classification. In *Papers from the AAAI-98 Workshop on Text Categorization*, pages 41–48, Madison, WI, July 1998.
25. Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision (IJCV-2004)*, 60(1):63–86, 2004.
26. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.
27. Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM-2000)*, pages 86–93, 2000.
28. Naoko Nitta, Noboru Babaguchi, and Tadahiro 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*, page 4718, Washington, DC, USA, 2000. IEEE Computer Society.
29. John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Peter J. Bartlett, Bernhard Schölkopf, Dale Schuurmans, and Alex J. Smola, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Boston, 1999.
30. Ariadna Quattoni, Micheal Collins, and Trevor Darrell. Learning visual representations using images with captions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-2007)*, pages 1–8. IEEE CS Press, June 2007.
31. Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05)*, volume 1, pages 29–36, Washington, DC, USA, 2005. IEEE Computer Society.
32. Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04)*, volume 3, pages 32–36, Washington, DC, USA, 2004. IEEE Computer Society.
33. Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM.
34. Jinqiao Wang, Lingyu Duan, Lei Xu, Hanqing Lu, and Jesse S. Jin. Tv ad video categorization with probabilistic latent concept learning. In *Multimedia Information Retrieval*, pages 217–226, 2007.
35. Yang Wang, Payam Sabzmeydani, and Greg Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. In *2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation*, 2007.
36. 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.