Exemplar-SVMs for Action Recognition

Christina Peterson
University of Central Florida
clp8199@knights.ucf.edu

Rui Hou
University of Central Florida
houray@gmail.com

Abstract

This goal of this paper is to introduce a method for action recognition that significantly reduces the labeling process. The method involves training a separate linear support vector machine (SVM) classifier for each selected exemplar and combining the scores to form mid-level features. Our approach is trained and tested on the UCF Sports Action data set. The accuracies achieved by the combined Exemplar-SVMs method approaches the best reported accuracies on the UCF Sports Action data set using a reduced amount of labeled data.

1. Introduction

Action recognition is an important problem in computer vision with applications including surveillance systems, content-based video retrieval, and human-computer interaction [10]. Solving action recognition by machine learning requires a large amount of labeled data to overcome the large intra-class variation regarding motion patterns and scene appearance. This paper addresses this problem by using an Exemplar-SVM, which can be defined as a linear SVM classifier trained on one positive instance. In order to generalize an action class, a set of Exemplar-SVMs are selected which serve as a good representation of the action. The methods proposed in this paper are based on the work of Malisiewicz et al. [5]. While their approach focuses on object detection, our approach focuses on action recognition. We obtain rich feature vectors for an action by using the base features described in [2]. We extend the feature vector to include space-time interest points (STIP) so that Histogram of Optical Flow (HOF) data is included. Our method uses the Exemplar-SVM scores as mid-level features similar to the work of [10].

2. Background

The concept of an Exemplar-SVM was first introduced by Malisiewicz et al. [5]. An Exemplar-SVM is a linear SVM classifier defined by one positive instance and millions of negative instances. Malisiewicz et al. [5] demonstrate that the performance of an ensemble of Exemplar-SVMs on the PASCAL VOC 2007 detection task is on par with the much more complex latent part-based model of Felzenszwalb et al. [3], at only a modest computational cost increase.

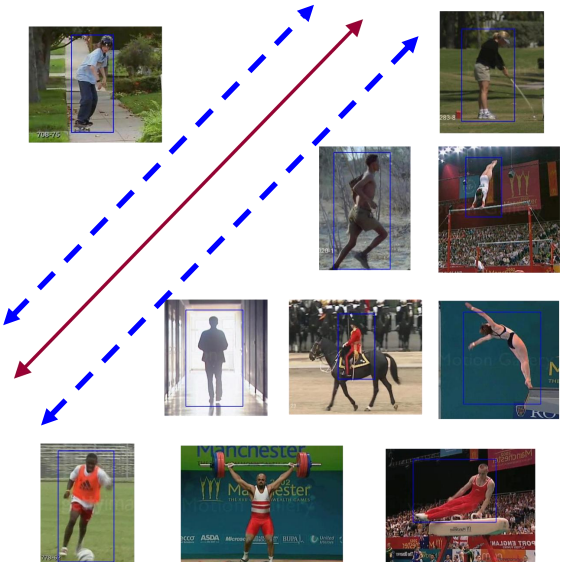


Figure 1. Exemplar-SVM Training Using One Positive Sample

Each exemplar can be represented by low-level features. The performance of each exemplar is dependent on which types of features are selected. Farhadi et al. [2] propose a method for object detection which focuses on categorizing objects by their semantic and discriminative attributes. A set of base features are defined in order to describe several visual aspects. Color and texture features are used to describe materials. Histogram of gradient (HOG) is used to describe parts. Edges are used to describe shapes. Our method adopts color, texture, and HOG features as presented by [2]. However, because our problem is focused on action recognition, our method also includes histogram of optical flow (HOF) to describe motion.

In order to localize an action, each exemplar includes all

low-level feature information of the action without the extraneous background data. Tian et al. [9] approach action detection by generalizing deformable part models from 2D images to 3D spatiotemporal volumes. A root filter is defined which captures the overall information of the action cycle. The root filter can be described as a bounding box which encloses the action cycle. The bounding box is divided spatially into 3x3 cells. The temporal division is determined such that each stage of the model contains 5 - 10 frames. Features are collected for each cell of the bounding box.

An exemplar is very specific to the instance it was trained on because it is based on only one positive sample. Therefore, a single exemplar is insufficient to generalize an entire action class. However, a series of exemplars has been shown to achieve good generalizations. Tran and Torresani [10] introduce Movement Exemplar-SVMs (MEXSVMs) used as a mid-level representation to achieve efficient recognition of actions in videos. Each MEXSVM is a linear classifier trained to recognize a specific instance within an action class. Their method learns N_a MEXSVMs, where the subvolume of each MEXSVM is run at N_s different scales over an input video. Max-pooling of the classifier scores is performed within N_p spatial-temporal pyramid volumes. This produces a mid-level feature vector of N_a x N_s x N_p dimensions.

3. Approach

3.1. Bag of Words

Low-level features were collected for all videos within the data set. These features include color, texture, and space-time interest points (STIP). Color and texture features are computed for each pixel. STIP features are computed using dense sampling. All features are quantized by k-means using 128 cluster centers for color, 256 cluster centers for texture, and 1000 cluster centers for STIP.

A bounding box is placed over each action within a video which encloses the action cycle and eliminates extra background information. The bounding box is spatially divided into 3x3 cells. The bounding box is temporally divided based on the time interval that STIP data was collected. For each feature, a normalized histogram of cluster center membership is computed for every cell within the bounding box. A final normalized histogram of the entire bounding box is computed for each feature. The feature vector for the bounding box is constructed by concatenating the histograms together which results in (128 + 256 + 1000) x (3x3 + 1) = 13,840 dimensions.

3.2. Exemplar-SVM Training

From each action class of the data set, a set of 3 - 5 videos are selected as the exemplar set. The videos within the exemplar set must be a good representation of the action class. There must be high chance that a randomly selected video within the action class will have similar features to at least one video within the exemplar set. The exemplar set for each action class is used for training the data set.



Figure 2. Exemplars Selected for Swing-Bench Action Class

Based on the work of [5], a linear classifier $f(x) = w^T\phi(x) + b$ is trained for each selected exemplar by minimizing the following objective function:

$$\min_{w,b} ||w||^2 \; + \; C_1 h(w^T\phi(x_E) + b) \\ + \; C_2 \sum_{x \in N_E} h(-w^T\phi(x) - b) \tag{1}$$

The exemplar feature vector is the positive sample and feature vectors from other actions are the negative samples. The background volume of each video within the training set is also included in the negative samples.

3.3. Calibration

The Exemplar-SVMs are trained independently of each other. This causes the raw output scores of each Exemplar-SVM to be incompatible. Platt [6] proposes a method for extracting the probabilities from SVM outputs by fitting the raw SVM scores to a sigmoid function. In order to perform calibration, each Exemplar-SVM is run on a validation set using a sliding volume method. An overlap score between the resulting detections and ground-truth is calculated by dividing the area of overlap by the area of the ground-truth bounding box. A detection with an overlap score greater than 0.5 is treated as positive. A detection with an overlap score lower than 0.2 is treated as negative. The raw SVM scores and corresponding overlap scores are plotted on a graph and a sigmoid function is fit to the data. The sigmoid function computes the calibrated detection score given a detection x and learned sigmoid parameters α and β as shown in equation (2).

$$f(x, \alpha, \beta) = \frac{1}{1 + e^{-\alpha(x-\beta)}} \tag{2}$$

3.4. Calibrated Exemplar-SVM Scores as Mid-Level Features

The calibrated Exemplar-SVM scores are used as mid-level features similar to the methods of [10]. Each video in the validation set is evaluated by N_a Exemplar-SVMs for

Method	Accuracy (%)	Diving	Golfing	Kicking	Lifting	Horse Riding	Running	Skate-Boarding	Swing-Bench	High-Swing	Walking
Rodriguez et al. [7]	69.2	68	61	66	66	75	74	73	-	-	-
Yeffet and Wolf [12]	79.3	100	61	65	67	67	69	92	-	-	86
Le et al. [4]	86.5	100	77.8	80	100	66.7	69.2	83.3	100	100	90.9
Wu et al. [11]	91.3	100	88	100	100	67	93	84	95	93	91
Sadanand et al. [8]	95.0	100	100	100	83	100	91	92	100	89	86
Standard Multiclass-SVM	90.2	100	50	71.4	100	100	100	100	100	100	50
Exemplar-SVMs (Ours)	94.4	96.6	87.5	91.7	100	83.3	100	98.1	97.9	94.4	82.5

Table 1. Recognition Accuracies of UCF Sports Data Set

Action	Di	Go	Ki	Li	Ho	Ru	Sk	Sb	Hs	Wa
Diving	96.6	1.9								1.5
Golf		87.5	8.3		4.2					
Kick		8.3	91.7							
Lift				100						
Horse-Ride					83.3		16.7			
Run						100				
Skateboard		1.9					98.1			
Swing-Bench								97.9	2.1	
High-Swing								2.8	94.4	2.8
Walk		3.3	3.3				10.8			82.5

Table 2. Confusion Matrix for Combined Exemplar-SVM

the data set. The raw svm scores and learned sigmoid parameters for each Exemplar-SVM are passed to the sigmoid function to obtain the calibrated scores. All of the calibrated scores for each Exemplar-SVM are concatenated together to form a mid-level feature vector of N_a dimensions for the validation video.

3.5. Multi-Class SVM Training Based on Calibrated Exemplar-SVM Scores

A Multi-Class SVM is trained on the mid-level feature vectors for the validation set. Based on the work of Chang and Lin [1], given the training samples $x_i \in R^n, i = 1, ..., l$, in two classes and an indicator vector $y \in R^l$ such that $y_i \in \{1,\text{-}1\}$, C-Support Vector Classification (C-SVC) solves the following primal optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w \; + \; C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, ..., l, \end{aligned} \tag{3}$$

where $\phi(x_i)$ maps x_i into a higher-dimensional space and $C > 0$ is the regularization parameter. The multi-class SVM is designed to classify localized actions of a data set.

4. Experimental Evaluation

Training, validation, and testing were done on the UCF Sports Action data set. A minimal exemplar set was selected so that there is a high probability that a randomly selected test video will have similar features to a video in the exemplar set. Training and testing are performed using LIBSVM [1]. A linear kernel type was used for training. Each Exemplar-SVM was trained using a two class method. The ground-truth of the exemplar video was used as the positive sample and the ground-truth of other actions in the training set were used as negative samples. Background volumes of videos within the training set were also used as negative samples.

The remaining videos disjoint from the exemplar set were split into two groups to be used for validation and testing. For best results, the videos of the validation set should resemble the videos of the testing set according to the base features. A two-fold cross validation method was used where validation was performed on the first group and testing performed on the second group. The groups were then swapped where validation was performed on the second group and testing was performed on the first group. The two-fold cross validation was performed a total of three times where a different split between validation and testing groups was chosen each time. The results were then averaged for each test case.

A sliding volume is performed on the validation set. The size of the sliding volume is based on the size of the ground-truth bounding box for the respective video. The Exemplar-SVM raw score is obtained for all volumes within the video. Calibration for each Exemplar-SVM is performed and the raw SVM scores of the validation set are converted to probabilities using the sigmoid function and learned sigmoid parameters. A multi-class SVM is trained on the mid-level feature vectors constructed from the calibrated Exemplar-SVM scores. A linear kernel type was used to train the multi-class SVM.

A sliding volume is then performed on the test set. The Exemplar-SVM raw scores are obtained for the volumes of the test set and converted to probabilities. The mid-level feature vectors are constructed for the positive detections and predicted by the multi-class SVM. The accuracy results are shown in Table 1. A confusion matrix for the Combined Exemplar-SVM is shown in Table 2. The accuracies of our method approach the best reported accuracies for this data set. However, our reported accuracies naturally inherited a bias in our favor given that we trained and tested within the same data set.

5. Conclusion

We have presented a method for action recognition which addresses the problem of intra-class variation regarding motion patterns and scene appearance. Our method approaches the best reported accuracies on the UCF Sports Action data set with the benefit of a reduced amount of required labeled data. This approach is recommended for recognition of actions with limited labeled training data.

Acknowledgment

The authors would like to thank Dr. Mubarak Shah and Dr. Niels da Vitoria Lobo for their constructive criticism and advice concerning this project.

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] H. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009.

[3] P. F. Felzenszwalb, R. B. Girschick, D. McCallester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[4] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatiotemporal features for action recognition with independent subspace analysis. *CVPR*, 2011.

[5] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. *ICCV*, 2011.

[6] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classiers*, 1999.

[7] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio- temporal maximum average correlation height lter for action recognition. *CVPR*, 2008.

[8] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. *CVPR*, 2012.

[9] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. *CVPR*, 2013.

[10] D. Tran and L. Torresani. Mexsvms: Mid- level features for scalable action recognition. *Dartmouth Computer Science Technical Report TR2013-726*, 2013.

[11] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. *CVPR*, 2011.

[12] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. *ICCV*, 2009.