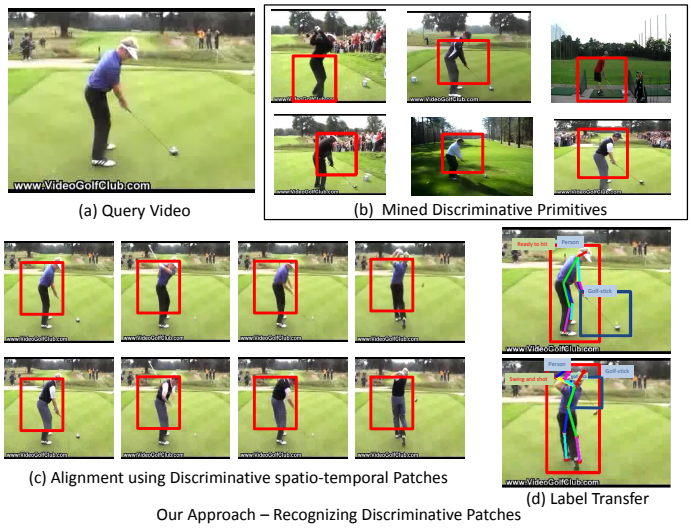
Representing Videos using Mid-level Discriminative Patches

Arpit Jain, Abhinav Gupta, Mikel Rodriguez, Larry S. Davis
ajain@umd.edu,abhinavg@cs.cmu.edu,mdrodriguez@mitre.org,lsd@cs.umd.edu

Abstract

We propose a new representation for videos based on mid-level discriminative patches. Previous approaches required these mid-level patches to have semantic meaning but in our approach they can correspond to a primitive human action, a semantic object, or perhaps a random but informative spatio-temporal patch in the video. Our only criteria for selecting these patches is their discriminative and representative properties. We propose an approach to automatically mine these patches from the training data. We also show that these patches establish strong correspondences with the test videos which can be used for label transfer. Furthermore, these patches can be used as a discriminative vocabulary for action classification where we get state-of-the-art performance on UCF50 and Olympics Sports datasets.



1. Introduction

We represent videos in terms of discriminative spatio-temporal patches rather than global feature vectors or a set of semantic objects. These spatio-temporal patches can be a primitive human action, a semantic object, or perhaps a random but informative spatio-temporal patch in the video. They are determined by their discriminative properties and their ability to establish correspondences with videos from similar classes. We automatically mine these discriminative patches from training data consisting of hundreds of videos. Figure 1(b) shows some of the mined discriminative patches for the "golf swing" class. We show how these mined patches can act as a discriminative vocabulary for action classification and demonstrate state-of-the-art performance on the Olympics Sports dataset [6] and the UCF-50 dataset[1]. But, more importantly, we demonstrate how these patches can be used to establish strong correspondence between spatio-temporal patches in training and test videos. We can use these correspondences to align the videos and perform tasks such as object localization, finer-level action detection etc. using label transfer techniques (refer Figure 1 (c) and (d)). Specifically, we present an integer-programming (IP) framework for selecting the set of mutually-consistent correspondences that best explains the classification of a video from a particular category. We then use these correspondences for representing the structure of a test video.

[1] http://server.cs.ucf.edu/vision/public_html/data.html

2. Approach

Given a set of training videos, we first find discriminative spatio-temporal patches which are representative of each action class. These patches satisfy two conditions: 1) they occur frequently within a class; 2) they are distinct from patches in other classes. The challenge is that the space of potential spatio-temporal patches is extremely large given that these patches can occur over a range of scales. And, the overwhelming majority of video patches are uninteresting, consisting of background clutter (track, grass, sky etc).

We address these issues by using an exemplar based clustering approach [1] which avoids partitioning the entire feature space. Every spatio-temporal patch is considered as a possible cluster center and we determine whether or not a discriminative cluster for some action class can be formed around that patch. We use the exemplar-SVM (e-SVM) approach of Malisiewicz et al. [5] to learn a discriminative distance metric for each cluster. However, learning an e-SVM for every spatio-temporal patch in the training dataset is computationally infeasible; instead, we use motion based

sampling to generate a set of initial cluster centers and then use simple nearest neighbor(NN) verification to prune candidates.

We pick few hundreds candidates from each class after NN pruning. We learn e-SVM on each of these patches and use it to form clusters by retrieving similar patches from the training and validation partitions. Finally, we re-rank these clusters based on their **(a)Appearance Consistency:** consistency score is computed by summing up the SVM detection scores of the top (10) detection scores from the validation partition.**(b) Purity:** To represent the purity/discriminativeness of each cluster, we use tf-idf scores.

We first evaluate our discriminative patches for action classification task. We select the top n e-SVM detectors from each class and apply them in a sliding cuboid fashion to a test video. We divide each video into a hierarchical 2-level grid and spatially max-pool the SVM scores in each cell to obtain the feature vector for a video [4]. We then learn a discriminative SVM classifier for each class using the features extracted on the training videos.

However our goal is to understand videos in finer details using these discriminative patches. Figure 1(c) shows how these patches establish strong correspondences between training and test videos. These correspondences can be used to perform a variety of other tasks such as object localization, finer-level action detection, etc. using simple label transfer (Refer Figure 1(d)). We have hundreds of discriminative patches in our vocabulary and many of them fire on the test video. Our goal is to select a subset of these patches which explains the video in the most coherent manner. We formulate this problem as a subset selection problem and solve it using IP framework.

Suppose, we have a vocabulary of size N. We have N possible candidate detections ($\{D_1, D_2, \ldots, D_N\}$) to select from. For each detection D_i, we associate a binary variable x_i which represents whether or not the detection of e-SVM i is selected. We first classify the video using our approach. If our inferred action class is l, then our goal is to select the x_i such that the cost function $\mathcal{J}_l$ is minimized.

$$\mathcal{J}_l = -\sum_i A_i x_i - w_1 \sum_i C_{li} x_i + w_2 \sum_{i,j} x_i P_{ij} x_j \quad (1)$$

where A_i is the zero centered normalized svm score for detection i, C_{li} is the class-consistency term which selects detections consistent with action class l and P_{ij} is the penalty term which encourages selection of detections which are consistent and discourages simultaneous detections from e-SVMs which are less likely to occur together. For optimizing the cost function, we use the IPFP algorithm [3].

3. Experiment

We demonstrate the effectiveness of our representation for the task of action classification and establishing correspondence. We use two benchmark action recognition datasets for experimental evaluation: UCF-50 and Olympics Sports Dataset [7]. For UCF-50 dataset, we train on 20 groups and test on 5 groups. We get an improvement of 3.32% over action-bank [8] (group-wise). Table 1 shows class-wise performance. On Olympics sport dataset, we get improvement of 2.6% over [2]. We also evalute how well our discriminative patches establish correspondences and align the videos. We manually labeled 50 discriminative patches per class with extra annotations such as objects of interaction (e.g, weights in clean-and-jerk), person bounding boxes and human poses. For 50 randomly sampled transfers, more than 50% of the transferred joints are within 15 pixels of the ground-truth joint locations. We also evaluated the localization performance of our algorithm for humans in the videos based on correspondence. We achieved 84.11% accuracy in localizing persons using 50% overlap criteria.

BaseballPitch	37.5	BasketBall	60.0	BenchPress	94.0
Biking	40.0	Billards	100	BreastStroke	100
CleanAndJerk	70.0	Diving	71.0	Drumming	47.1
Fencing	77.3	GolfSwing	69.0	HighJump	44.4
HorseRace	88.0	HorseRiding	90.4	HulaHoop	30.8
JavelinThrow	36.4	JugglingBalls	13.6	JumpRope	25.0
JumpingJack	88.0	Kayaking	57.1	Lunges	30.8
MilitaryParade	57.7	Mixing	40.7	Nunchucks	0
PizzaTossing	20.8	PlayingGuitar	60.0	PlayingPiano	95.0
PlayingTabla	65.2	PlayingViolin	50.0	PoleVault	84.4
PommelHorse	73.1	Pullup	45.8	Punch	90.3
PushUps	59.1	RockClimbingIndoor	77.8	RopeClimbing	60.0
Rowing	70.4	SalsaSpin	57.1	SkateBoarding	61.5
Skiing	59.3	Skijet	85.0	SoccerJuggling	53.3
Swing	64.0	TaiChi	55.0	TennisSwing	51.5
ThrowDiscus	71.0	TrampolineJumping	75.0	VolleyballSpiking	91.3
WalkingWithDog	47.8	YoYo	62.5		

Table 1. Quantitative Evaluation on UCF50 dataset

References

[1] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 2012. 1

[2] A. Gaidon, Z. Harchaoui, and C. Schmid. Recognizing activities with cluster-trees of tracklets. In *BMVC*, 2012. 2

[3] M. Leordeanu, M. Hebert, and R. Sukthankar. An integer projected fixed point method for graph matching and map inference. In *NIPS*, 2009. 2

[4] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 2

[5] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1

[6] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 1

[7] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 2008. 2

[8] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012. 2