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

How should a video be represented? We propose a new representation for videos based on mid-level discriminative spatio-temporal patches. These spatio-temporal patches might correspond to a primitive human action, a semantic object, or perhaps a random but informative spatio-temporal patch in the video. What defines these spatio-temporal patches is their discriminative and representative properties. We automatically mine these patches from hundreds of training videos and experimentally demonstrate that these patches establish correspondence across videos and align the videos for label transfer techniques. Furthermore, these patches can be used as a discriminative vocabulary for action classification where they demonstrate state-of-the-art performance on UCF50 and Olympics datasets.

1. Introduction

Consider the video visualized as a spatio-temporal volume in Figure 1a. What does it mean to understand this video and how might we achieve such an understanding? Currently, the most common answer to this question involves recognizing the particular event or action that occurs in the video. For the video shown in the figure it would simply be "clean and jerk" (Figure 1b). But this level of description does not address issues such as the temporal extent of the action [27]. It typically uses only a global feature-based representation to predict the class of action. We additionally would like to determine structural properties of the video such as the time instant when the person picks up the weight or where the weights are located.

We want to understand actions at a finer level, both spatially and temporally. Instead of representing videos globally by a single feature vector, we need to decompose them into their relevant "bits and pieces". This could be addressed by modeling videos in terms of their constituent semantic actions and objects [10, 30, 4]. The general framework would be to first probabilistically detect objects (e.g, weights, poles, people) and primitive actions (e.g, bending and lifting). These probabilistic detections could then be combined using Bayesian networks to build a consistent and coherent interpretation such as a storyline [11] (Figure 1c). So, the semantic objects and actions form primitives for representation of videos. However, recent research in object and action recognition has shown that current computational models for identifying semantic entities are not robust enough to serve as a basis for video analysis [7]. Therefore, such approaches have, for the most part, only been applied to restricted and structured domains such as baseball [11] and office scenes [30].

Following recent work on discriminative patch-based representation [2, 31], we represent videos in terms of discriminative spatio-temporal patches rather than global feature vectors or a set of semantic entities. These spatio-temporal patches might correspond to a primitive human action, a semantic object, human-object pair 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(d)(left) shows some of the mined discriminative patches for the "weightlifting" 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 [23] 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 this correspondence to align the videos and perform tasks such as object localization, finer-level action detection etc. using label transfer techniques [6, 22]. Specifically, we present an integer-programming 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. Figure 2 shows an example of how aligned videos (shown in Figure 1(d)(right)) are used to localize humans and objects, detect finer action categories and estimate human poses.

2. Prior Work

Prior approaches to video representation can be roughly divided into three broad categories. The first and earliest represent actions using global spatio-temporal templates,

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

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.332

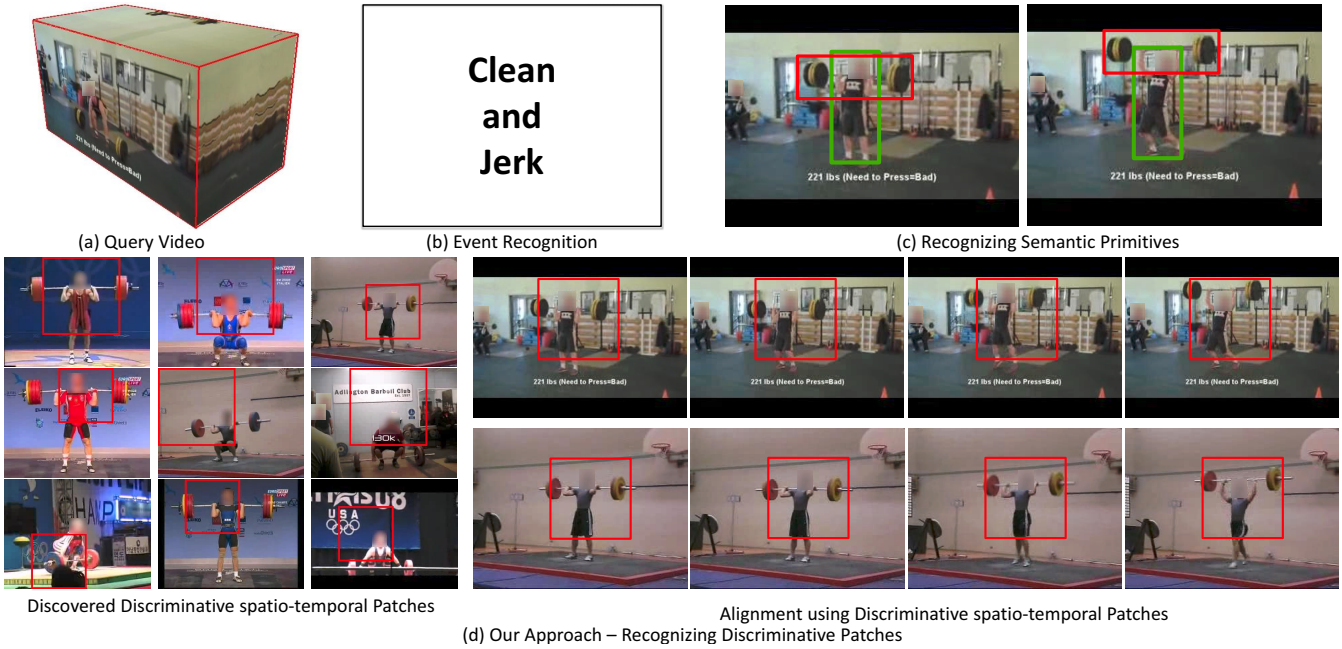

Figure 1. Given a query video (a), one can represent it using global feature vector and use it for action classification (b). Another possible representation is to use constituent semantic entities (c) and use object/action detectors for understanding. Instead, we propose a mid-level representation for videos (d). Our approach discovers representative and discriminative spatio-temporal patches for a given action class (d-left). These patches are then used to establishing correspondence followed by alignment (d-right).

such as motion history [1] and spatiotemporal shapes [9] .

The second class of approaches is based on bag of features models [17, 24, 28], where sparse spatio-temporal interest points, dense interest points [32], page-rank features [21], or discriminative class-specific features [15], are computed as part of a bag of words representation on local features. Typically these representations are most appropriate for classification; they are not well-suited as action detectors or for establishing correspondence.

The third class of approaches is structural and decomposes videos into constituent parts. These parts typically correspond to semantic entities such as humans and objects [10, 34, 14]. While these approaches attempt to develop a rich representation and learn the structure of the videos in terms of constituent objects, one of their inherent drawbacks is that they are highly dependent on the success of object and action detection algorithms. Therefore, such approaches have not been used for "data in the wild". A more recent approach is based on using discriminative spatio-temporal patches rather than semantic entities [12, 26]. For example, [26] uses manually selected spatio-temporal patches to create a dictionary of discriminative patches for each action class. These patches are then correlated with test video patches and a new feature vector is created using pooling. There are several issues here: 1) What is the criteria for selecting spatio-temporal patches to create the dictionary? 2) How many patches are needed to capture all the variations in the data? Motivated by work in object recognition [7], recent approaches have attempted to decompose an action or event into a set of discriminative "parts" or spatio-temporal "patches" designed to capture the local spatio-temporal structure of the data [33, 23]. However, these approaches still focus on the problem of classification and cannot establish strong correspondence or explain why a video is classified as a member of certain class.

Our approach is similar in spirit to work on poselets in object recognition [2]. The key idea is that instead of using semantic parts/constituents, videos are represented in terms of discriminative spatio-temporal patches that can establish correspondences across videos. However, learning poselets requires key-point annotation, which is very tedious for videos. Furthermore, for general videos it is not even clear what should actually be labeled. Recent approaches have tried to circumvent the key point annotation problem by using manually-labeled discriminative regions [16] or objectness criteria [29] to create candidate discriminative regions. We do not use any priors (such as objectness) to select discriminative patches; rather we let the data select the patches of appropriate scale and location. We build upon the recent work of Singh et al. [31] and extract "video poselets" from just action labels. However, the huge scale of our data (videos) precludes direct use of [31]. Therefore, we propose an efficient exemplar-SVM based procedure to cluster the data without partitioning the whole space. We also propose a principled IP framework for selecting correspondences at the test time. Also, note that our approach is different from multiple instance learning [25] since we do not

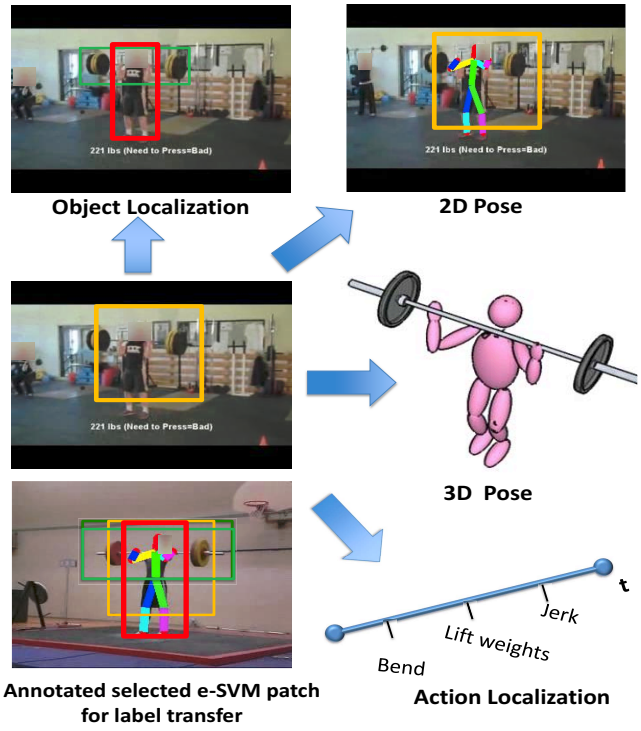

Figure 2. Strong alignment allows us to richly annotate test videos using a simple label transfer technique.

assume that there exists a consistent spatio-temporal patch across positive examples (positive instance in the bag); instead we want to extract multiple discriminative patches per action class depending on the style in which an action is performed.

3. Mining Discriminative Patches

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).

One approach would be to follow the bag-of-words paradigm: sample a few thousand patches, perform k-means clustering to find representative clusters and then rank these clusters based on membership in different action classes. However, this has two major drawbacks: **(a) High-Dimensional Distance Metric:** K-means uses standard distance metrics such as Euclidean or normalized cross-correlation. These standard distance metrics do not work well in high-dimensional spaces(In our case, we use HOG3D [13] to represent each spatio-temporal patch and the dimensionality of the feature space is 1600). For example, Figure 3 shows a query patch (left) and similar patches retrieved using Euclidean distance (right). The Euclidean distance fails to retrieve visually similar patches. Instead, we learn a discriminative distance metric to retrieve similar patches and, hence, representative clusters. **(b) Partitioning:** Standard clustering algorithms partition the entire feature space. Every data point is assigned to one of the clusters during the clustering procedure. However, in many cases, assigning cluster memberships to rare background patches is hard. Due to the forced clustering they significantly diminish the purity of good clusters to which they are assigned.

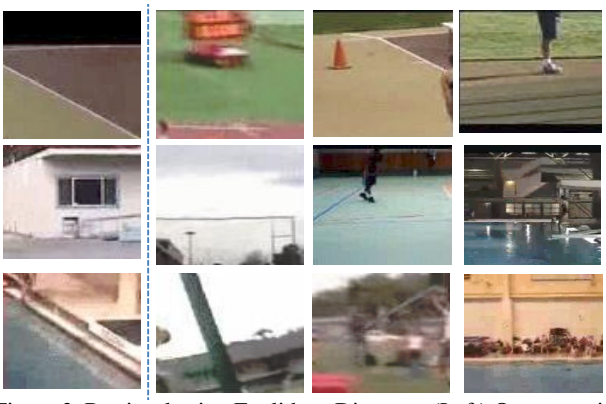

Figure 3. Retrieval using Euclidean Distance. (Left) Query spatio-temporal patch. (Right) Retrieval using euclidean distance metric.

We address these issues by using an exemplar-based clustering approach [5] 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. [22] 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 verification to prune candidates. The following section presents the details of this algorithm.

3.1. Approach

Available training data is partitioned into training and validation sets. The training partition is used to learn a discriminative distance metric and form clusters and the validation partition is used to rank the clusters based on representativeness. We sample a few hundred patches from each video in the training partition as candidates. We bias the sampling to avoid background patches - patches with uniform or no motion should be rejected.

However, learning an e-SVM for all the sampled patches is still computationally infeasible (Assuming 50 training videos per class and 200 sampled patches, we have approx-

imately 10K candidate patches per class). Therefore, we perform pruning using a simple nearest-neighbor approach. For each spatio-temporal patch, we determine its k (=20, typically) nearest neighbors in the training partition. We score each patch based on how many nearest neighbors are within class as opposed to the number out of class. Based on this ranking, we select a few hundred patches per action class and use the e-SVM to learn patch-specific discriminative distance metrics. These e-SVMs are then used to form clusters by retrieving similar patches from the training and validation partitions. Finally, we re-rank the clusters using the approach described next.

3.2. Ranking

Our goal is to select a smaller dictionary (set of representative patches) from the candidate patches for each class. Our criteria for ranking consists of two terms: **(a) Appearance Consistency:** We use the SVM classifier confidence as the measure of appearance consistency. The 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: the ratio of how many patches it retrieves from videos of the same action class to the number of patches retrieved from videos of different classes.

All patches are ranked using a linear combination of the two scores. Figure 4 shows a set of top-ranked discriminative spatio-temporal patches for different classes selected by this approach. As the figure shows, our spatio-temporal patches are quite representative of various actions. For example, for discus-throw, our approach extracts the patch corresponding to the turning motion before the throw (see 1st column and 2nd row) and for pull-ups it extracts the up-down motion of the body (see 1st column, 1st row). As expected, our discriminative patches are not always semantically meaningful. Also, notice how our clusters exhibit good visual correspondences, which can be exploited for label transfer. To further demonstrate that our patches are quite representative and capture the essence of actions, we extracted spatio-temporal patches that exemplify "Gangnam" style dance. We use 30 gangnam dance step youtube videos as our positive set and 340 random videos as a negative set. Figure 5 shows the top discriminative patches selected which indeed represent the dance steps associated with gangnam-dance.

4. Analyzing Videos

4.1. Action Classification

We first evaluate our discriminative patches for action classification. We select the top n e-SVM detectors from each class and apply them in a sliding cuboid fashion to a test video. Similar to object-bank [20], we construct a feature vector based on the results of the e-SVMs. 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. We then learn a discriminative SVM classifier for each class using the features extracted on the training videos.

4.2. Beyond Classification: Explanation via Discriminative Patches

We now discuss how we can use detections of discriminative patches for establishing correspondences between training and test videos. Once a strong correspondence is established and the videos are aligned, we can perform a variety of other tasks such as object localization, finer-level action detection, etc. using simple label transfer (see Figure 6).

Our vocabulary consists of hundreds of discriminative patches; many of the corresponding e-SVMs fire on any given test video. This raises a question: which detections to select for establishing correspondence. One could simply use the SVM scores and select the top-scoring detections. However, individual e-SVM detections can lead to bad correspondences. Therefore, we employ a context-dependent approach to jointly select the e-SVM detections across a video. We formulate a global cost function for selection of these detections and use relaxed integer programming to optimize and select the detections.

Context-dependent Patch Selection: For simplicity, we consider the top detection of each e-SVM as a candidate detection for selection, although the approach can be extended to allow multiple (but bounded) numbers of firings of any patch. Therefore, if 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. Our goal is to select the subset of detections which: (a) have high activation score (SVM score); (b) are consistent with the classified action; (c) are mutually consistent. We first classify the video using the methodology described in Section 4.1. 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. We explain each term in detail:

- **Appearance term:** A_i is the e-SVM score for patch i. This term encourages selection of patches with high e-SVM scores.
- **Class Consistency:** C_{li} is the class consistency term. This term promotes selection of certain e-SVMs over

Figure 4. Examples of highly ranked discriminative spatio-temporal patches.

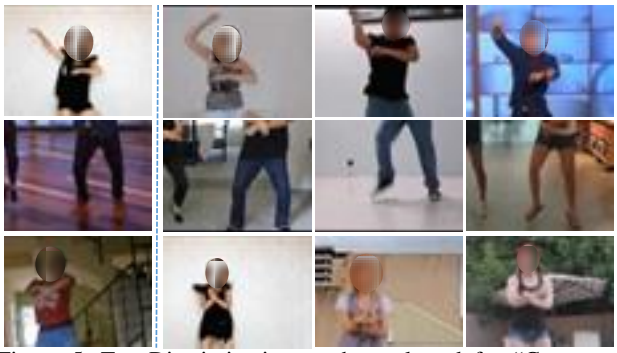

Figure 5. Top Discriminative patches selected for "Gangnam" style.

others given the action class. For example, for the weightlifting class it prefers selection of the patches with man and bar with vertical motion. We learn C_l from the training data by counting the number of times that an e-SVM fires for each class.

- **Penalty term:** P_{ij} is the penalty term for selecting a pair of detections together. We penalize if: 1) e-SVMs i and j do not fire frequently together in the training data; 2) the e-SVMs i and j are trained from different action classes. We compute co-occurrence statistics of pairs of eSVMs on the training data to compute the penalty.

Optimization: The objective function results in an Integer Program which is an NP-hard problem. For optimizing the cost function, we use the IPFP algorithm proposed in [19]. IPFP algorithm is very efficient and the optimization converges in 5-10 iterations. IPFP solves quadratic optimization functions of the form:

$X^* = \text{argmax}(X_n^T M X_n) \quad \text{s.t. } 0 \leq X_n \leq 1$

To employ IPFP, we transform the cost function to the above form through the following substitution: $X_n = \begin{pmatrix} 1 \\ X \end{pmatrix}$ and $M = \begin{pmatrix} 1 & \frac{(A+C)^T}{2} \\ \frac{(A+C)}{2} & -P \end{pmatrix}$.

The solution obtained by the IPFP algorithm is generally binary, but if the output is not binary then we threshold at 0.5 to binarize it. The set of patches which maximizes this cost function is then used for label transfer and to infer finer details of the underlying action.

5. Experimental Evaluation

We demonstrate the effectiveness of our representation for the task of action classification and establishing correspondence. We will also show how correspondence between training and test videos can be used for label transfer and to construct detailed descriptions of videos.

Datasets: We use two benchmark action recognition datasets for experimental evaluation: UCF-50 and Olympics Sports Dataset [24]. We use UCF-50 to qualitatively evaluate how discriminative patches can be used to establish correspondences and transfer labels from training to test videos. We manually annotated the videos in 13 of these classes with annotations including the bound-

ing boxes of objects and humans (manually annotating the whole dataset would have required too much human effort). We also performed a sensitivity analysis (w.r.t. to vocabulary size) on this subset of UCF-50 dataset. Quantitatively, we evaluate the performance of our approach on action classification on the UCF-50 and the complete Olympics dataset.

Implementation Details: Our current implementation considers only cuboid patches, and takes patches at scales ranging from 120x120x50 to the entire video. Patches are represented with HOG3D features (4x4x5 cells with 20 discrete orientations). Thus, the resulting feature has 4x4x5x20 = 1600 dimensions. At the initial step, we sample 200 spatio-temporal patches per video. The nearest neighbor step selects 500 patches per class for which e-SVMs are learned. We finally select a vocabulary of 80 e-SVMs per class. During exemplar learning, we use a soft-margin SVM with C fixed to 0.1. The SVM parameters for classification are selected through cross validation.

5.1. Classification Results

UCF Dataset: The UCF50 dataset can be evaluated in two ways: videowise and groupwise. We tested on the more difficult task of groupwise classification guaranteeing that the backgrounds and actors between the training and test sets are disjoint. We train on 20 groups and test on 5 groups. We evaluate performance by counting the number of videos correctly classified out of the total number of videos in each class. Table 1 shows performance of our algorithm compared to the action bank approach [26] on the 13 class subset (run with same test-train set as our approach) and a bag-of-words approach as a baseline. We also evaluated performance with respect to vocabulary size. Table 4 shows the performance variation with the number of e-SVM patches trained per class. Finally, we evaluate action classification for all 50 classes in UCF (group-wise) and get an improvement of 3.32% over action-bank.

Olympics Dataset: We follow the same experimental setting for splitting the data into test-train and employ the same evaluation scheme (mAP) as used in [23]. Tables 2 and 3 show the performance of our approach versus previous approaches.

Action Class	BoW(baseline)	[26]	Ours
Basketball	20.00	**53.84**	50.00
Clean and Jerk	40.00	85.00	**95.65**
Diving	58.06	**78.79**	61.29
Golf Swing	54.84	**90.32**	75.86
High Jump	12.90	38.46	**55.56**
Javeline Throw	29.03	45.83	**50.00**
Mixing	12.90	42.85	**55.56**
PoleVault	65.62	60.60	**84.37**
Pull Up	48.88	**91.67**	75.00
Push Ups	40.63	85.00	**86.36**
Tennis Swing	**51.51**	44.12	48.48
Throw Discus	63.64	75.00	**87.10**
Volleyball Spiking	24.24	43.48	**90.90**
Mean Classification	40.17	64.23	**70.47**

Table 1. Classification performance of our algorithm compared to Action Bank [26] in groupwise division of dataset

Sport Class	[23]	[18]	[3]	Ours
High-jump	68.9	52.4	75.8	**84.94**
Long-jump	74.8	66.8	78.6	**84.64**
Triple-jump	52.3	36.1	69.7	**83.29**
Pole-vault	82.0	47.8	85.5	**84.67**
Gymnastics-Vault	86.1	88.6	**89.4**	82.58
Shot-put	62.1	56.2	65.9	**83.55**
Snatch	69.2	41.8	72.1	**83.47**
Clean-jerk	84.1	83.2	86.2	**86.64**
Javelin-throw	74.6	61.1	77.8	**84.75**
Hammer-throw	77.5	65.1	79.4	**86.40**
Discus-Throw	58.5	37.4	62.2	**86.66**
Diving-platform-10m	87.2	**91.5**	89.9	86.51
Diving-springboard-3m	77.2	80.7	82.2	**86.44**
Basketball-layup	77.9	75.8	79.7	**88.60**
Bowling	72.7	66.7	78.7	**88.27**
Tennis-serve	49.1	39.6	63.8	**83.37**

Table 2. Quantitative Evaluation on Olympics Sports Dataset. Mean results are shown in the next table.

5.2. Correspondence and Label Transfer

We now demonstrate how our discriminative patches can be used to establish correspondence and align the videos. Figure 6 shows a few examples of alignment using the detections selected by our framework. It can be seen that our spatio-temporal patches are insensitive to background changes and establish strong alignment. We also use the aligned videos to generate annotations of test videos by simple label-transfer technique. 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. After aligning the videos we transfer these annotations to the new test videos.

Figure 6 shows the transfer of annotations. These examples show how strong correspondence can allow us to perform tasks such as object detection, pose estimation and predicting temporal extent. For example, detecting the golf club in the golf-swing case is extremely difficult because the golf club occupies very few pixels in the video. But our strong alignment via motion allows us to transfer the bounding box of the golf-club to the test video. Similarly, estimating human poses for golf-swing and discus throw would be extremely difficult. But again, using our discriminative spatio-temporal patches we just align the videos using motion and appearance and then transfer the poses from the training videos to test videos. We also did an informal evaluation of our pose transfer. For 50 randomly sampled trans-

Approach	mAP
Niebles et. al. [23]	71.1
Laptev et. al. [18]	62.0
William et. al. [3]	77.3
Adrien et. al. [8]	82.7
Ours	85.3

Table 3. Comparison on Olympics Dataset

Patches per class	mAP
20	62.52
30	65.25
50	67.57
80	70.47
100	70.17

Table 4. Effect of Vocabulary Size on UCF13

fers, 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.

Conclusion: We proposed a new representation for videos based on spatio-temporal patches. We automatically mine these patches from hundreds of training videos using exemplar-based clustering approach. We have also shown how these patches can be used to obtain strong correspondence and align the videos for transferring annotations. Furthermore, these patches can be used as a vocabulary to achieve state of the art results for action classification.

Acknowledgement: This research is partially supported by ONR N000141010766 and Google. It is also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. *D*isclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 2001. 2

[2] L. Bourdev and J. Malik. Poselets:body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 1, 2

[3] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011. 6, 7

[4] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 1

[5] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 2012. 3

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003. 1

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1, 2

[8] A. Gaidon, Z. Harchaoui, and C. Schmid. Recognizing activities with cluster-trees of tracklets. In *BMVC*, 2012. 7

[9] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 2

[10] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 2009. 1, 2

[11] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 1

[12] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007. 2

[13] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 3

[14] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013. 2

[15] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010. 2

[16] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 2

[17] I. Laptev. On space-time interest points. *IJCV*, 2005. 2

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 6, 7

[19] M. Leordeanu, M. Hebert, and R. Sukthankar. An integer projected fixed point method for graph matching and map inference. In *NIPS*, 2009. 5

[20] 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. 4

[21] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009. 2

[22] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 3

[23] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 1, 2, 6, 7

[24] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 2008. 2, 5

[25] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 2

[26] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012. 2, 6

[27] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010. 1

[28] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, 2007. 2

[29] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: An application to weakly supervised action classification. In *ECCV*, 2012. 2

[30] Z. Si, M. Pei, and S. Zhu. Unsupervised learning of event and-or grammar and semantics from video. In *ICCV*, 2011. 1

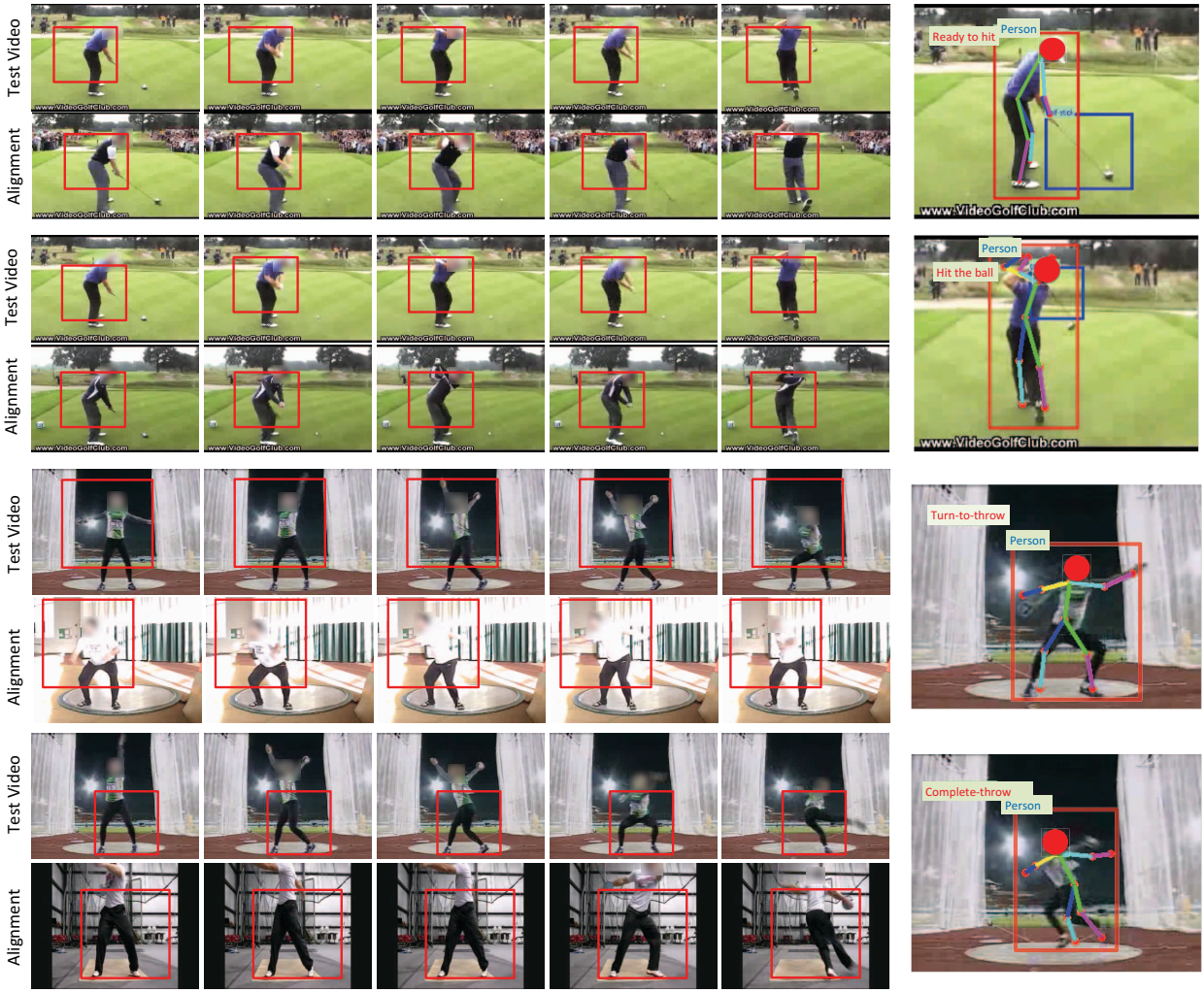

Figure 6. Rich Annotations using Label Transfer: Discriminative patches can be used to align test video with training videos. Once videos are aligned, annotations such as object bounding boxes and human poses can be obtained by simple label transfer.

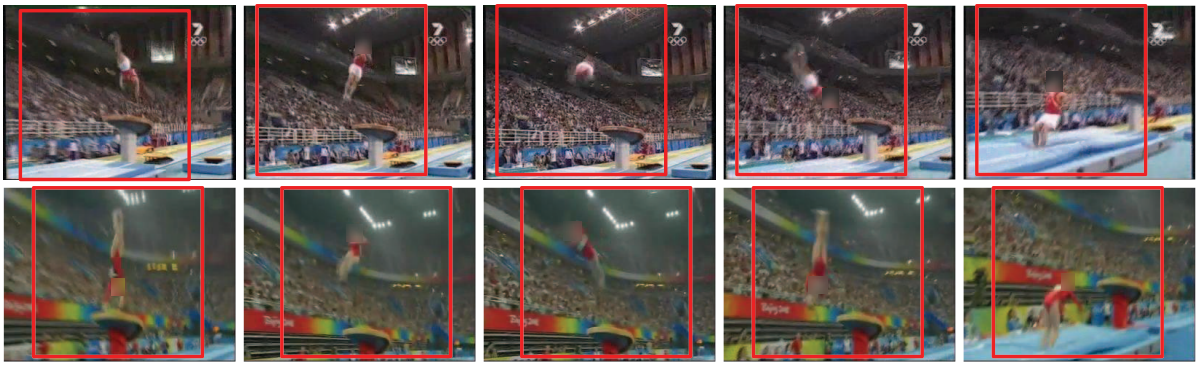

Figure 7. Example alignment for the Olympics Dataset.

[31] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 1, 2

[32] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2

[33] Y. Wang and G. Mori. Hidden part models for human action recognition:probabilistic versus max margin. *PAMI*, 2011. 2

[34] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2