Final Version Dated: 31 May 2012
IEEE Xplore DOI: http://doi.ieeecomputersociety.org/10.1109/CVPR.2012.6247806
All data and scripts to reproduce the full set of results are available at http://www.cse.buffalo.edu/~jcorso/r/actionbank

# Action Bank: A High-Level Representation of Activity in Video

Sreemananth Sadanand and Jason J. Corso
Computer Science and Engineering, SUNY at Buffalo
{sreemana,jcorso}@buffalo.edu

## Abstract

*Activity recognition in video is dominated by low- and mid-level features, and while demonstrably capable, by nature, these features carry little semantic meaning. Inspired by the recent object bank approach to image representation, we present* Action Bank*, a new high-level representation of video. Action bank is comprised of many individual action detectors sampled broadly in semantic space as well as viewpoint space. Our representation is constructed to be semantically rich and even when paired with simple linear SVM classifiers is capable of highly discriminative performance. We have tested action bank on four major activity recognition benchmarks. In all cases, our performance is better than the state of the art, namely 98.2% on KTH (better by 3.3%), 95.0% on UCF Sports (better by 3.7%), 57.9% on UCF50 (baseline is 47.9%), and 26.9% on HMDB51 (baseline is 23.2%). Furthermore, when we analyze the classifiers, we find strong transfer of semantics from the constituent action detectors to the bank classifier.*

## 1. Introduction

Human motion and activity is extremely complex; automatically inferring activity from video in a robust manner that would lead to a rich high-level understanding of video remains a challenge despite the great energy the vision community has invested in it. The most promising current approaches are primarily based on low- and mid-level features such as local space-time features [18], dense point trajectories [36], and dense 3D gradient histograms [15] to name a few; these methods have demonstrated capability on realistic data sets like UCF Sports [30]. But, they are, by nature, limited in the amount of motion semantics they can capture being strictly low-level, which often yields a representation with inadequate discriminative power for larger, more complex data sets. For example, on the 50-class UCF50 data set [1], the HOG/HOF method [18, 37] achieves 47.9% accuracy (as reported in [17]) whereas it achieves 85.6% on the smaller 9-class UCF Sports data set [30]. Other methods that seek a more semantically rich and discriminative representation have focused on object and scene semantics [13] or human pose, e.g., [2, 29], which itself is challenging and unsolved.

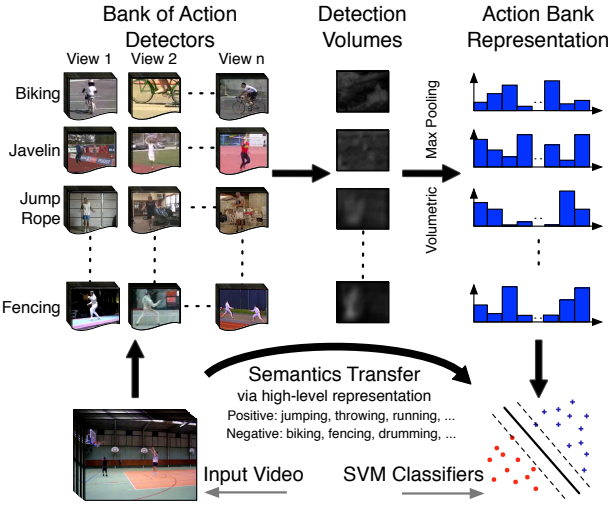


Figure 1. Action bank is a high-level representation for video activity recognition. Inspired by the object bank method [22], action bank stores a large set of individual action detectors (at varying scales and viewpoints). The action bank representation is a concatenation of volumetric max-pooled detection volume features from each detector. This high-level representation transfers the semantics from the bank entries through to the output (Section 2.5).

In this paper, we propose a new high-level representation of human action in video that we call Action Bank. Inspired by the Object Bank method [22], action bank explores how a large set of action detectors, which ultimately act like the bases of a high-dimensional "action-space," combined with a simple linear classifier can form the basis of a semantically-rich representation for activity recognition and other video understanding challenges (Figure 1 shows an overview). The individual action detectors in the action bank are based on an adaptation of the recent action spotting framework [6] and hence template-based; despite the great amount of research on action recognition, few methods are available that localize action in the video


Figure 2. A montage of entries in the action bank, 36 of the 205 in the bank. Each entry in the bank is a single template video example (see Section 2.3 for details on how these are used as detectors). The columns depict different types of actions, e.g., a baseball pitcher, boxing, etc. and the rows indicate different examples for that action. Examples are selected to roughly sample the action's variation in viewpoint and time (but each is a different video/scene, i.e., this is not a multiview requirement). Faces are redacted for presentation only.

as a detector must. Individual detectors in the bank capture example actions, such as "running-left" and "biking-away," and are run at multiple scales over the input video (many examples of detectors in action bank are shown in Figure 2). The outputs of detectors are transformed into a feature vector by volumetric max-pooling. Although the resulting vector is high-dimensional, we test an SVM classifier that is able to enforce sparsity among its representation, in a manner similar to object bank.

Although there has been some work on mid- and high-level representations for video recognition and retrieval [13], to the best of our knowledge it has exclusively been focused on object and scene-level semantics, such as face detection. Our work hence breaks new ground in establishing a high-level representation built atop individual action detectors. We show that this high-level representation of human activity is capable of being the basis of a powerful activity recognition method (Section 3), achieving better than state-of-the-art accuracies on every major activity recognition benchmark attempted, including 98.2% on KTH [33], 95.0% on UCF Sports [30], 57.9% on the full UCF50 [1], and 26.9% on HMDB51 [17]. Furthermore, action bank also transfers the semantics of the individual action detectors through to the final classifier (Section 2.5).

**Action Detection vs. Action Recognition.** Despite the great emphasis on action recognition in the past decades in the computer vision literature—activity recognition is a core component of comprehensive image and video understanding—there is comparatively little work on action detection. Emphasis has been on classifying whether an action is present or absent in a given video, rather than detecting where and when in the video the action may be happening. We will next survey some of the related literature on action recognition, and then we will discuss template-based methods on which we base action bank because they essentially do *recognition by detection*.

As mentioned in the previous section, perhaps the most studied and successful approaches thus far in activity recognition are based on bag of features (dense or sparse) models. Introduced by [18], sparse space-time interest points and subsequent methods, such as local trinary patterns [39], dense interest points [37], page-rank features [24], and discriminative class-specific features [16], typically compute a bag of words representation on local features and sometimes local context features that is used for classification. A recent trend has been to use densely rather than sparsely sampled features for better performance, e.g., [38] who achieve scores as high as 91.3% and 94.5% overall accuracy on the UCF Sports [30] and KTH [33] data sets, respectively. In summary, although promising, these methods are predominantly global recognition methods and are not well-suited as action detectors.

A second class of methods rely upon an implicit ability to find and process the human before recognizing the action. For example, [12] develop a space-time shape representation of the human motion from a segmented silhouette. Joint-keyed trajectories [2] and pose-based methods [29] involve localizing and tracking human body parts prior to modeling and performing action recognition. Obviously this class of methods is better suited to localizing action, but the challenge of localizing and tracking humans and human pose has limited their adoption.

Action bank uses template-based methods because they naturally do recognition by detection (frequently through simple convolution) and do not require complex human localization, tracking or pose. Early template-based action recognition methods use optical flow-based representation [7, 8, 28]. Later methods avoid the explicit computation of optical flow due to its complexity and limitations: Bobick and Davis [3] compute a two-vector of motion presence and

recency at each pixel. The Action MACH method [30] fuses multiple examples into a single template via Clifford algebras on vector-fields of spatiotemporal regularity flow. Derpanis et al. [6] propose "action spotting," a template representation that also forgoes explicit motion computation. The representation is based on oriented space-time energy, e.g., leftward motion and flicker motion, and is invariant to (spatial) object appearance, and efficiently computed by separable convolutions [5]. Action bank uses this spotting approach for its individual detectors due to its capability (invariant to appearance changes), simplicity, and efficiency.

## 2. The Action Bank Representation of Videos

Action bank represents a video as the collected output of many action detectors that each produce a correlation volume. Although, in spirit, action bank is closely related to object bank [22], in practice, we have found the action problem to be distinct from the object problem, as we now explain. We use a template-based action detector (Section 2.3) as the primary element of action bank. The detector is invariant to changes in appearance, but we have needed to carefully infuse robustness/invariance to scale, viewpoint, and tempo. To account for changes in scale, we run the detectors at multiple scales, similar to object bank. But, to account for viewpoint and tempo changes, we sample variations of them for each action. Figure 2 gives many good examples of this sampling; take the left column—baseball pitcher—which we sample from the front, left-side, right-side and rear, whereas in the second-column we sample both one and two-person boxing in quite different settings.

We select actions from standard data sets and provide full details on which actions and how they are selected in Section 3. Once constructed, we use the same action bank for our entire paper. The nature of the representation warrants inquiry regarding "how big" and "how diverse" the action bank needs to be. These are complex questions to answer theoretically and instead we carry out a thorough empirical investigation on these questions in Section 3.4 to ultimately find bigger is better but may be over-kill depending on the size of the action class-space.

### 2.1. The Action Bank Feature Vector

For a given action bank with $N_a$ detectors, each action detector is run at $N_s$ scales (spatiotemporal) to yield $N_a \times N_s$ correlation volumes. We adapt the max-pooling method in [20] to the volumetric case (see Figure 3) and take three levels in the octree. For each action-scale pair, this amounts to $1^3 + 2^3 + 4^3$ or a 73 dimension vector. The total length of the action bank feature vector is hence $N_a \times N_s \times 73$.

### 2.2. Training and Classifying with Action Bank

We use a standard SVM classifier on the action bank feature vector. Although structural risk minimization is used in object bank [22], we have not found it to outperform the standard hinge loss with $L2$ regularization (Section 3.3). Being a template-based method, there is actually no training of the individual bank detectors. Presently, we manually select which detector templates are in the bank. In the future, we foresee an automatic process of building the action bank by selecting best-case templates from among those possible. Nevertheless, we find that only a small subset of the actions in the bank have a nonzero weight in the SVM classifier. We liken this fact to a *feature selection* process in which the bank detectors serve as a large feature pool and the training process selects a subset of them, hence mitigating the manual-selection of the individual bank templates being a limiting factor. At present the manual approach has led to a powerful action bank that can perform significantly better than current methods on activity recognition benchmarks.

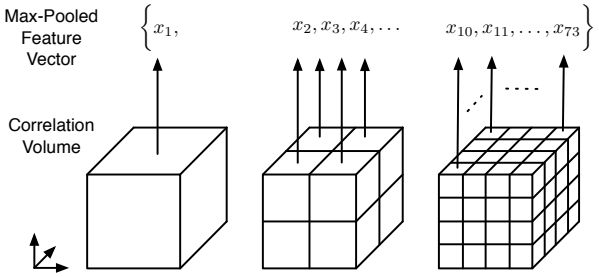

Figure 3. Volumetric max-pooling extracts a spatiotemporal feature vector from the correlation output of each action detector.

### 2.3. Action Templates as Bank Detectors

Action bank allows a great deal of flexibility in choosing what kind of action detectors are used; indeed different types of action detectors can be used concurrently. In our implementation, we use the recent "action spotting" detector [6] due to its desirable properties of invariance to appearance variation, evident capability in localizing actions from a single template, efficiency (is implementable as a set of separable convolutions [5]), and natural interpretation as a decomposition of the video into space-time energies like leftward motion and flicker. We do make a modification of the original action spotting method to increase its sensitivity to the action and its efficiency; in this section, we explain the action spotting method and our variation of it.

**Actions as composition of energies along spatiotemporal orientations.** An action can be considered as a conglomeration of motion energies in different spatiotemporal orientations. Consider that motion at a point is captured as a combination of energies along different space-time orientations at that point, when suitably decomposed. These decomposed motion energies are a low-level action representation and the basis of the action spotting method [6].

A spatiotemporal orientation decomposition is realized using broadly tuned 3D Gaussian third derivative filters,

$G_{3_{\hat{\theta}}}(x)$, with the unit vector $\hat{\theta}$ capturing the 3D direction of the filter symmetry axis and $x$ denoting space-time position. The responses of the image data to this filter are pointwise squared and summed over a space-time neighbourhood $\Omega$ to give a pointwise energy measurement

$$E_{\hat{\theta}}(x) = \sum_{x \in \Omega} (G_{3_{\hat{\theta}}} * I)^2 \quad . \tag{1}$$

A basis-set of four third-order filters is then computed according to conventional steerable filters [9]:

$$\hat{\theta}_i = \cos\left(\frac{\pi i}{4}\right)\hat{\theta}_a(\hat{n}) + \sin\left(\frac{\pi i}{4}\right)\hat{\theta}_b(\hat{n}) \tag{2}$$

where $\hat{\theta}_a(\hat{n}) = \hat{n} \times \hat{e_x}/\|\hat{n} \times \hat{e_x}\|, \hat{\theta}_b(\hat{n}) = \hat{n} \times \hat{\theta}_a(\hat{n})$, $\hat{e}$ is the unit vector along the spatial x axis in the Fourier domain and $0 \leq i \leq 3$. And this basis-set makes it plausible to compute the energy along any frequency domain plane—spatiotemporal orientation—with normal $\hat{n}$ by a simple sum $E_{\hat{n}}(x) = \sum_{i=0}^{3} E_{\hat{\theta}_i}(x)$ with $\hat{\theta}(i)$ as one of the four directions calculated according to (2).

For our action bank detector, we define seven raw spatiotemporal energies (via different $\hat{n}$): static $E_s$, leftward $E_l$, rightward $E_r$, upward $E_u$, downward $E_d$, flicker $E_f$, and lack of structure $E_o$ (which is computed as a function of the other six and peaks when none of the other six has strong energy). Finally, we have experimentally found that these seven energies do not always sufficiently discriminate action from common background. So, we observe that lack of structure $E_o$ and static $E_s$ are disassociated with any action and use their signal to separate the salient energy from each of the other five energies, yielding a five-dimensional *pure* orientation energy representation: $\overline{E}_i = E_i - E_o - E_s \quad \forall i \in \{f, l, r, u, d\}$. Finally, the five pure energies are normalized such that the energy at each voxel over the five channels sums to one.

**Template matching.** Following [6], we use a standard Bhattacharya coefficient $m(\cdot)$ when correlating the template $T$ with a query video $V$:

$$M(x) = \sum_{u} m\left(V(x-u), T(u)\right) \tag{3}$$

where $u$ ranges over the spatiotemporal support of the template volume and $M(\cdot)$ is the output correlation volume; the correlation is implemented in the frequency domain for efficiency. Conveniently, the Bhattacharya coefficient bounds the correlation values between 0 and 1, with 0 indicating a complete mismatch and 1 indicating a complete match, which gives an intuitive interpretation for the correlation volume that is used in volumetric max-pooling.

## 2.4. Neurophysiological Evidence

Although action bank is not a *biologically inspired* method, there is indeed evidence in the neurophysiological literature to justify the proposed method of building and applying a bank of action detectors in the manner we do. There is neurophysiological evidence that mammalian brains have an action bank-like representation for human motion. Perrett et al. [27] discovered that neurons in the superior temporal sulcus of the macaque monkey brain were selective to certain types of mammalian motion, such as head rotation. Early research in human motion perception has also suggested that humans recognize complex activities as the composition of simpler canonical motion categories, such as that of a swinging pendulum [14]. Finally and most significantly, other neurophysiological research, e.g., [10], suggests that view-specific representations are constructed in the visual pathway. For instance, recognition of certain point-light motions degrades with the angle of rotation away from the learned viewpoint. These view-specific exemplars (templates) of action are exactly what comprise our action bank (see, for example, Figure 2).

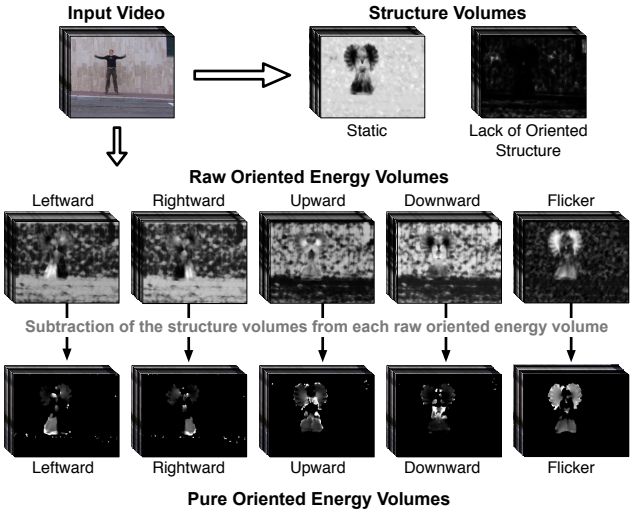


Figure 4. A schematic of the spatiotemporal orientation energy representation that is used for the action detectors in action bank. A video is decomposed into seven canonical space-time energies: leftward, rightward, upward, downward, flicker (very rapid changes), static, and lack of oriented structure; the last two are not associated with motion and are hence used to modulate the other five (their energies are subtracted from the raw oriented energies) to improve the discriminative power of the representation. The resulting five energies form our appearance-invariant template.

## 2.5. Looking Inside Action Bank

Given the high-level nature of the action bank representation, we investigate the question of whether the semantics of the representation have actually transferred into the classifiers. For example, does the classifier learned for a running activity pay more attention to the *running-like* entries in the bank than it does other entries, such as *spinning-like*? We perform our analysis by plotting the dominant (posi-

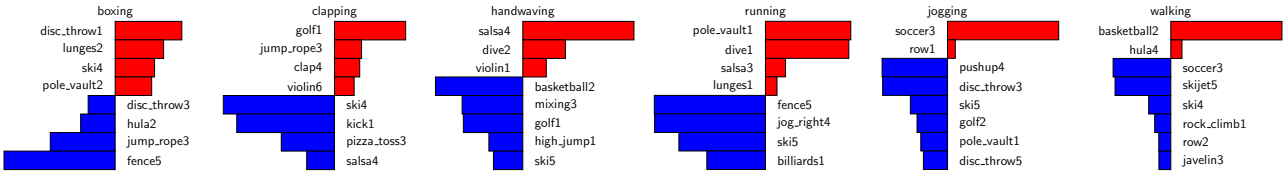


Figure 5. Relative contribution of the dominant positive and negative bank entries for each one-vs-all SVM on the KTH data set. The action class is named at the top of each bar-chart; red (blue) bars are positive (negative) values in the SVM vector. The number on bank entry names denotes which example in the bank (recall that each action in the bank has 3–6 different examples). Note the frequent semantically meaningful entries; for example, "clapping" incorporates a "clap" bank entry and "running" has a "jog" bank entry in its negative set.

tive and negative) weights of each one-vs-all SVM weight vector; see Figure 5 where we have plotted these dominant weights for the six classes in KTH. We select the top-four (when available; in red; these are positive weights) and bottom-four (or more when needed; in blue; these are negative weights).

Close inspection of which bank entries are dominating verifies that some semantics are transferred into the classifiers; but, indeed, some unexpected transfer happens as well. Encouraging semantics-transfers[1] include but are not limited to positive "clap4" selected for "clapping" and even "violin6" selected for "clapping" (we suspect due to the back and forth motion of playing the violin); positive "soccer3" for "jogging" (the soccer entries are essentially jogging and kicking combined) and negative "jog_right4" for "running". An unexpected semantics-transfers is positive "pole_vault4" and "ski4" for "boxing" (for which we have no good explanation but do note similar behavior has been observed in object bank [22]).

## 3. Experimental Evaluation

Our experiments demonstrate the proposed approach for activity recognition for a wide variety of activity categories in realistic video and on a larger scale than has been typical in recent papers. Sections 3.1 and 3.2 contain the comparative evaluation using KTH [33], UCF Sports [30], UCF50 [1], and HMDB51 [17] data sets. In all four cases, action bank outperforms all known methods in the literature, and in some cases by a significant margin. In addition to raw performance experiments, we analyze how the action bank representation works with different classifiers (Section 3.3), and with banks of different sizes (Section 3.4). Finally, section 3.5 describes the computational cost of action bank.

**Building the bank.** The action bank used for all experiments consists of 205 template actions collected from all 50 action classes in UCF50 [1] and all six action classes from KTH [33]. We use three to six action templates from the same action but each being shot from different views and scales, but note this is not "multiview" action recognition as these templates are of different people in different videos. When selecting the templates, we have sought to roughly sample the different viewpoints and temporal scales; we have constructed only one bank and it is used in all of the experiments, without any manual tweaking or optimization. The action templates have an average spatial resolution of approximately $50 \times 120$ pixels and a temporal length of $40 - 50$ frames (examples are in Figure 2); each template is cropped spatially to cover the extent of the human motion within it. No videos in the action bank are used in the experimental testing set.

[1] In these examples, "clap4", "violin6", "soccer3", "jog_right4", "pole_vault4', "ski4", "basketball2", and "hula4" are names of individual templates in our action bank.

### 3.1. Benchmark Action Recognition Datasets

We compare performance on the two standard action recognition benchmarks: KTH [33] and UCF Sports [30]. In these experiments, we run the action bank at two spatial scales (we do not modify the temporal scale). On KTH (Table 1 and Figure 6), we use the original splits from [33] with any testing videos in the bank removed. Our method, at 98.2%, outperforms all other methods, three of which share the current best performance of 94.5% [11, 16, 38]. Most of the previous methods reporting high scores are based on feature points and hence have quite a distinct character from action bank; following, it is interesting to note that we seem to confuse classes they understand and learn classes they confuse. We perfectly learn jogging and running whereas we confuse boxing and walking frequently; yet, other methods seem to most frequently confuse jogging and running.

| | hw | bx | wk | jg | cl | rn |
|---|---|---|---|---|---|---|
| handwaving | 1 | 0 | 0 | 0 | 0 | 0 |
| boxing | 0 | 0.92 | 0.08 | 0 | 0 | 0 |
| walking | 0 | 0.03 | 0.97 | 0 | 0 | 0 |
| jogging | 0 | 0 | 0 | 1 | 0 | 0 |
| clapping | 0 | 0 | 0 | 0 | 1 | 0 |
| running | 0 | 0 | 0 | 0 | 0 | 1 |

Figure 6. Confusion matrix for the KTH [33] data set.

We use a leave-one-out cross-validation strategy for UCF Sports as others have used in the community, but do not engage in horizontal flipping of the data as some have [16, 37, 38]. Again, our performance, at 95% accuracy, is better than all contemporary methods, who achieve at best

| Method | Accuracy (%) |
|---|---|
| Schüldt et al. [33] | 71.7 |
| Klaser et al. [15] | 84.3 |
| Savarese et al. [32] | 86.8 |
| Ryoo and Aggarwal [31] | 91.1 |
| Liu et al. [23] | 91.6 |
| Laptev et al. [19] | 91.8 |
| Bregonizo et al. [4] | 93.2 |
| Liu et al. [24] | 93.8 |
| Le et al. [21] | 93.9 |
| Liu and Shah [25] | 94.3 |
| Gilbert et al. [11] | 94.5 |
| Kovashka and Grauman [16] | 94.5 |
| Wu et al. [38] | 94.5 |
| Action Bank | **98.2** |

Table 1. Recognition accuracies on the KTH data set.

| Method | Accuracy (%) |
|---|---|
| Rodriguez et al. [30] | 69.2 |
| Yeffet and Wolf [39] | 79.3 |
| Varma and Babu [35] | 85.2 |
| Wang et al. [37] | 85.6 |
| Le et al. [21] | 86.5 |
| Kovashka and Grauman [16] | 87.3 |
| Wu et al. [38] | 91.3 |
| Action Bank | **95.0** |

Table 2. Recognition accuracies on the UCF Sports data set.

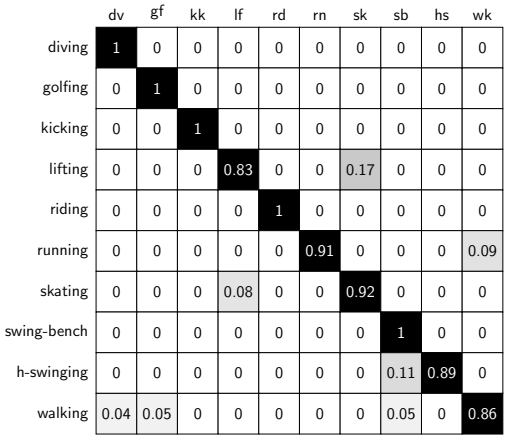


Figure 7. Confusion matrix for the UCF Sports [30] data set.

91.3% [38] (Table 2, Figure 7).

These two sets of results clearly demonstrate that action bank is a notable new representation for human activity in video and capable of robust recognition in realistic settings. However, these two benchmarks are small; next we move to more realistic benchmarks that are an order of magnitude larger in terms of classes and number of videos.

### 3.2. UCF50 and HMDB51: Larger Scale Tests

The action recognition data sets presented in Section 3.1 are all relatively small, ranging from $6-11$ classes and $150-599$ videos, and are unable to test the scalability of an action recognition system to more realistic scenarios. The UCF50 [1] data set, however, is better suited to test scalability: it has 50 classes and 6680 videos. We are aware of only [17] who have recently processed two methods through the UCF50 data set (and also released a new data set HMDB51 of similar size to UCF50).

We have processed the entire UCF50 data set through action bank (using a single scale for computational reasons) and ran it through 10-fold video-wise cross-validation (Figure 8 and Tables 3 and 4) and 5-fold group-wise cross-validation (Table 4) experiment. We first note that the baselines run on UCF50 by [17] perform lower than action bank. Our confusion matrix shows a dominating diagonal with no stand-out confusion among the classes; most frequently, skijet and rowing are inter-confused and yoyo is confused as nunchucks. Pizza-tossing is the worst performing class (46.1%) but its confusion is rather diffuse. The generalization from the data sets with much less classes to UCF50 is encouraging for our action bank representation. Furthermore, we have also run the experiment on the new HMDB51 data set (using the three-way splits in [17]) and find a similar relative performance of 26.9% to a baseline HOG/HOF performance of 20.2% (see Table 4).

### 3.3. Varying the Classifier

We have primarily used a standard SVM to train the classifiers in these experiments. However, given the emphasis on sparsity and structural risk minimization in the original

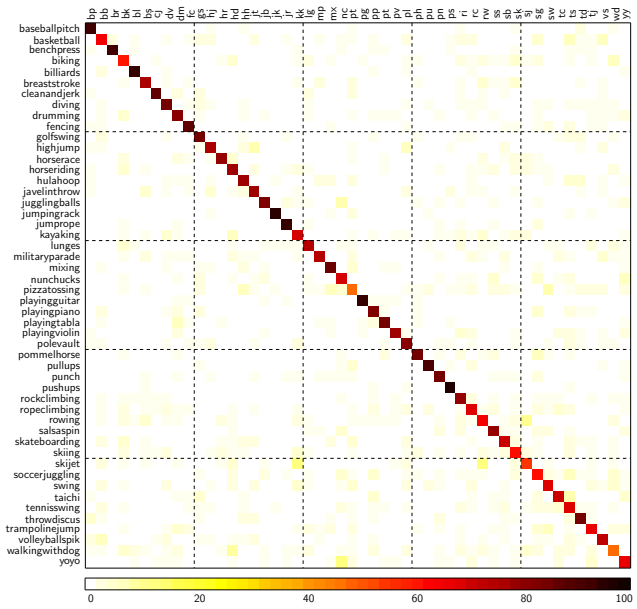


Figure 8. Confusion matrix for the UCF50 [1] data set. Numbers are not shown for clarity; the color legend is drawn at the bottom (please view in color). Table 3 shows the per-class accuracy in more detail.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pushups | 95.8 | golfswing | 83.5 | javelinthrow | 77.4 | ropeclimbing | 66.7 |
| jumpingrack | 93.7 | diving | 83.3 | playingviolin | 76.7 | yoyo | 66.4 |
| jumprope | 93.2 | throwdiscus | 83.1 | breaststroke | 75.8 | trampolinejump | 65.4 |
| playingguitar | 93.1 | pommelhorse | 82.9 | highjump | 75.7 | basketball | 64.2 |
| billiards | 92.6 | jugglingballs | 81.7 | militaryparade | 73.0 | rowing | 64.2 |
| baseballpitch | 91.9 | playingpiano | 81.1 | lunges | 73.0 | soccerjuggling | 61.7 |
| benchpress | 91.0 | polevault | 80.6 | volleyballspik | 72.4 | skiing | 60.5 |
| pullups | 88.9 | drumming | 80.0 | taichi | 71.1 | biking | 60.0 |
| cleanandjerk | 87.1 | salsaspin | 79.2 | skateboarding | 70.4 | skijet | 54.4 |
| fencing | 86.0 | horserace | 78.9 | nunchucks | 68.9 | walkingwithdog | 46.4 |
| mixing | 85.0 | rockclimb | 78.4 | kayaking | 68.3 | pizzatossing | 46.1 |
| punch | 84.0 | hulahoop | 77.7 | tennisswing | 68.0 | | |
| playingtabla | 83.8 | horseriding | 77.4 | swing | 67.5 | | |

Table 3. Per class (sorted) accuracy score for UCF50 [1].

| | Accuracy (%) | | |
|---|---|---|---|
| Method | UCF50-V | UCF50-G | HMBD51 |
| Gist [26] | - | 38.8 | 13.4 |
| Laptev et al. [18, 37] | - | 47.9 | 20.2 |
| C2 [17] | - | - | 23.2 |
| Action Bank | 76.4 | **57.9** | **26.9** |

Table 4. Comparing overall accuracy on UCF50 [1] and HMDB51 [17] (-V specifies video-wise CV, and -G group-wise CV). We have relied on the scores reported in [17] for the baseline Gist and HOG/HOF bag of words. We are not aware of other methods reporting results on UCF50 or HMDB51.

object bank work [22], we also test the performance of action bank when used as a representation for other classifiers, including a feature sparsity L1-regularized logistic regression SVM (LR1) and a random forest classifier (RF). We evaluated the LR1 on UCF50 and found the performance to drop to 71.1% on average, and we evaluated the RF on the KTH and UCF Sports data sets on which we found 96% and 87.9%, respectively. These efforts have demonstrated a degree of robustness inherent in the action bank classifier (accuracy does not drastically change) and that a standard SVM performs best, given our experiment conditions.

### 3.4. Varying the Size of the Bank

A driving factor in this work is the generality of the action bank to adapt to different video understanding settings: for a new setting, simply add more detectors to the bank. However, it is not given that a larger bank necessarily means better performance: the curse of dimensionality may counter this intuition. To assess it, we have conducted an experiment that varies the size of the bank from 5 detectors to the full 205 detectors. For each different size $k$, we run 150 iterations in which we randomly sample $k$ detectors from the full bank and construct a new bank. Then, we perform a full leave-one-out cross-validation on the UCF Sports data set. The results are reported in Figure 9, and as we expect, the bigger bank does indeed perform better, though not to the expected extent. With a bank of size 80 we are able to match the existing state-of-the-art score from [38], and with a bank of size 5, we achieve 84.7% accuracy, which is still surprisingly high.

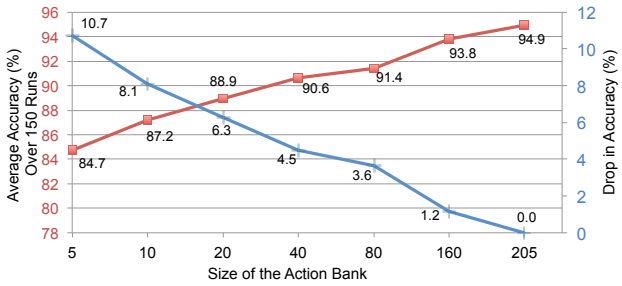


Figure 9. Experiment to analyze the effect of the bank size of the recognition accuracy. We vary the size of the bank from 5 elements to 205 by random sub-selection and then average the results over 150 runs. The red curve plots this average accuracy and the blue curve plots the drop in accuracy for each respective size of the bank with respect to the full bank. These results are on the UCF Sports data set. Clearly the strength of the method is maintained even for banks half as big as the one we have primarily used.

How can we explain this surprising stability for such a small bank? Consider the following interpretation of action bank: the bank is an embedding of the video into an action-space whose bases are the individual action detectors. A given activity is then described as a combination of these action detectors (seldom or never is a sole detector the only one firing in the bank). Recall Figure 5, the nonzero weights for boxing are "disc_throw", "lunges", etc. even though boxing templates are in the action bank. Furthermore, we also point out that, although we are not using a group sparsity regularizer in the SVM, we observe a gross group sparse behavior. For example, in the jogging and walking classes, only two entries in the bank have any positive weight and few have any negative weight. In most cases, $80-90\%$ of the bank entries are not selected; but across the classes, there is variation among which are selected. We believe this is because of the relative sparsity in our action bank detector outputs when adapted to yield pure spatiotemporal orientation energy (Section 2.3). With this evidence, the performance stability even with a very small bank is not very surprising: even in the case of a large bank, only a small subset of the actions in the bank are actually incorporated into the final classification.

### 3.5. Computational Cost

From a computational cost perspective, action bank is convolution. We have, of course, implemented our code to use FFT-based convolution but have otherwise not optimized it, nor are we using a GPU in our current implementation. On a 2.4GHz Linux workstation, the mean cpu-time used to process a video from UCF50 is 12210 seconds (204 minutes) with a range of $1560-121950$ seconds ($26-2032$ minutes or $0.4-34$ hours) and a median of 10414 seconds (173 minutes). As a basis of comparison, a typical bag of words with HOG3D method ranges between $150-300$ sec-

onds, a KLT tracker extracting and tracking sparse points ranges between 240-600 seconds, and a modern optical flow method [34] takes more than 24 hours on the same machine. If we parallelize the processing over 12 cpus by running the video over elements in the bank in parallel, we can drastically reduce the mean running time to 1158 seconds (19 minutes) with a range of $149 - 12102$ seconds ($2.5 - 202$ minutes) and a median of 1156 seconds (19 minutes).

## 4. Conclusion

We have presented Action Bank, a conceptually simple yet effectively powerful method for carrying out high-level activity recognition on a wide variety of realistic videos "in the wild." The method leverages on the fact that a large number of smaller action detectors, when pooled appropriately, can provide high-level semantically rich features that are superior to low-level features in discriminating videos—our results show a moderate to significant improvement from action bank on every major benchmark we have attempted, including both small and large-scale settings. Our method builds a high-level representation using the output of a large bank of individual, viewpoint-tuned action detectors. This high-level representation has rich applicability in a wide-variety of video understanding problems, and we have shown its capability on activity recognition in this paper.

We are investigating a few extensions to the current version of action bank. Namely, the process of building the action bank (i.e., selecting the templates) is prone to human error; we are looking into automatically building an action bank given a set of videos. We are also conscious of the increasing emphasis on real-time applications of computer vision systems; we are working on a method that will iteratively apply the action bank on streaming video by selectively sampling frames to compute based on an early coarse resolution computation.

**Source Code and Precomputed Results** The entire source code for action bank is available for download at `http://www.cse.buffalo.edu/~jcorso/r/actionbank`. We also make precomputed action bank feature vectors of many popular video data sets available at this site to foster easy use of the method.

**Acknowledgements** This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), the DARPA Mind's Eye program (W911NF-10-2-0062), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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, DARPA, ARO, NSF or the U.S. Government.

## References

[1] `http://server.cs.ucf.edu/~vision/data.html`.
[2] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007.
[3] A. Bobick and J. Davis. The Recognition of Human Movement Using Temporal Templates. *TPAMI*, 23(3):257–267, 2001.
[4] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR*, 2009.
[5] K. G. Derpanis and J. Gryn. Three-dimensional nth derivative of gaussian separable steerable filters. In *IEEE ICIP*, 2005.
[6] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *CVPR*, 2010.
[7] A. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.
[8] I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions. *TPAMI*, 19(7):757–763, 1997.
[9] W. Freeman and E. Adelson. The design and use of steerable filters. *TPAMI*, 13(9):891–906, 1991.
[10] M. A. Giese. *Neural model for the recognition of biological motion*, volume Dynamische Perzeption, pages 105–110. Infix Verlag, 2000.
[11] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *CVPR*, 2009.
[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29(12):2247–2253, 2007.
[13] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? *TMM*, 9(5):958–966, 2007.
[14] G. Johanssson. Visual-perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.
[15] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.
[16] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.
[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.
[18] I. Laptev. On space-time interest points. *IJCV*, 64(2/3):107–123, 2005.
[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[21] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.
[22] 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.
[23] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.
[24] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.
[25] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.
[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.
[27] D. I. Perrett, P. A. Smith, A. J. Mistlin, A. J. Chitty, A. S. Head, D. D. Potter, R. Broennimann, A. D. Milner, and M. A. Jeeves. Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report. *Behavioural Brain Research*, 16(2-3):153–170, 1985.
[28] R. Polana and R. Nelson. Low level recognition of human motion (or how to your main without funding his body parts). In *IEEE MNRAO Workshop*, 1994.
[29] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.
[30] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.
[31] M. S. Ryoo, M. Shah, and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *CVPR*, 2009.
[32] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *IEEE WMVC*, 2008.
[33] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.
[34] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.
[35] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *ICML*, 2009.
[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.
[37] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.
[38] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.
[39] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.