From Actemes to Action: A Strongly-supervised Representation for Detailed Action Understanding

Weiyu Zhang
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA, USA
zhweiyu@seas.upenn.edu

Menglong Zhu
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA, USA
menglong@cis.upenn.edu

Konstantinos G. Derpanis
Department of Computer Science
Ryerson University
Toronto, ON, Canada
kosta@scs.ryerson.ca

Abstract

This paper presents a novel approach for analyzing human actions in non-scripted, unconstrained video settings based on volumetric, x-y-t, patch classifiers, termed actemes. Unlike previous action-related work, the discovery of patch classifiers is posed as a strongly-supervised process. Specifically, keypoint labels (e.g., position) across spacetime are used in a data-driven training process to discover patches that are highly clustered in the spacetime keypoint configuration space. To support this process, a new human action dataset consisting of challenging consumer videos is introduced, where notably the action label, the 2D position of a set of keypoints and their visibilities are provided for each video frame. On a novel input video, each acteme is used in a sliding volume scheme to yield a set of sparse, non-overlapping detections. These detections provide the intermediate substrate for segmenting out the action. For action classification, the proposed representation shows significant improvement over state-of-the-art low-level features, while providing spatiotemporal localization as additional output. This output sheds further light into detailed action understanding.

1. Introduction

Human action classification ("*What action is present in the video?*") and detection ("*Where and when is a particular action performed in the video?*") are key tasks for understanding imagery. Addressing such tasks is rendered difficult by the wide variation of human appearance in unconstrained settings, due for instance to differences in clothing, body shape, lighting and camera viewpoint. Another challenge arises from the diversity in observed action dynamics due to performance nuances and varying camera capture settings, such as frame rate. Furthermore, different actions may have large common portions and thus can only be distinguished by transient and subtle differences, e.g., the lower body of a golf-swing and a baseball swing look the same. A key question for addressing these issues is the *selection of the visual primitives for representing actions*.

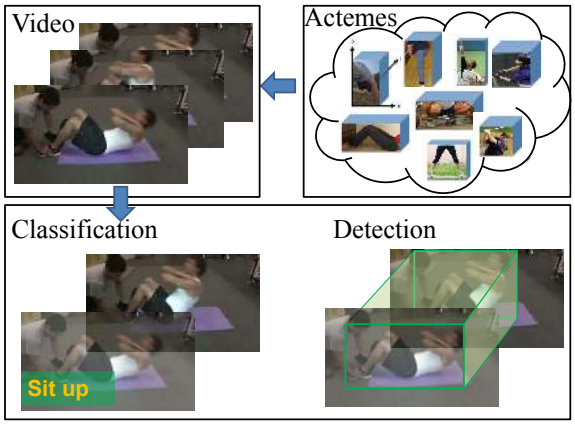

Figure 1. Overview of proposed approach. An input video is encoded by the activation of a set of discriminative spatiotemporal patch classifiers, **actemes**. Actemes can take on a variety of spatiotemporal extents. The activations of these mid-level primitives in the input imagery are used to classify videos into different action categories and spatiotemporally localize an action.

An intuitive representation for human actions is the underlying pose of limbs or joints across time. While impressive progress in pose estimation and tracking has been made [20], these tasks remain open problems in general unconstrained settings. Further, it is inconclusive whether such hard subproblems are necessary for addressing certain action-related problems, as opposed to the recovery of information from images in a more direct fashion [19].

Much of the state-of-the-art work in action classification and detection is based on relating image patterns in a more direct fashion to actions. A tremendous variety of visual spatiotemporal descriptors have appeared in the literature. One manner of organizing these features is in terms of their spatiotemporal extent/resolution.

At one extreme of the spectrum are local spatiotemporal descriptors [30]. Here, local spatiotemporal image patches are often encoded by a universal dictionary of "visual words", constructed by clustering a large number of training image patches in an unsupervised manner. Commonly, these descriptors are combined to define a global descriptor over the video, termed a "bag of visual words". In practice, most local patches capture generic oriented structures, such as bar and corner-like structures in the spatial [24] and spatiotemporal [18] domains. Thus, visual words in isolation generally convey generic information of the local spacetime pattern, such as the velocity and the presence of spatiotemporal discontinuities, rather than distinctive information of the action, including its spacetime location. Further, as typically employed, these features are not strictly grounded on actions but rather capture both the action and the scene context which may artificially increase performance on datasets where the actions are highly correlated with the scene context [17].

At the other end of the spectrum are spatiotemporal templates that represent the *entire* action. A variety of image measurements have been proposed to populate such templates, including optical flow and spatiotemporal orientations [6, 12, 4, 25]. Due to the inflexibility of these representations, an inordinate amount of training data is required to span the visual appearance space of an action.

In between the local and holistic representational extremes lie mid-level part representations that model moderate portions of the action. Here, parts have been proposed that capture a neighborhood of spacetime [13, 7, 29, 22], a spatial keyframe [26] or a temporal segment while considering the entire spatial extent of the scene [21, 9]. These representations attempt to balance the tradeoff between generality exhibited by small patches, e.g., visual words, and the specificity by large ones, e.g., holistic templates. The volumetric part/patch definitions proposed in the current work potentially span the spectrum. A data-driven approach informed by rich training data (i.e., keypoints across spacetime) is used to identify the optimal operating *points* between these two extremes.

Most closely related to the current work are poselets [1]. A poselet is an image patch classifier trained with the aid of annotated keypoint locations. Each classifier is selective for a particular salient *spatial* visual pattern corresponding to a portion of a human pose captured at a particular camera viewpoint. The composition of these parts capture the essence of an object. Poselets have been shown to be an effective mid-level representation for a variety of applications, including action recognition in *still* imagery [19].

Motivated by the success of poselets and other patch-based representations in static contexts, this work presents a set of volumetric patch detectors, termed *actemes*, that are each selective for a salient *spatiotemporal* visual pattern. An acteme can consist of a single frame outlining a keyframe of the action, a volume outlining the entire action or an intermediate patch size, capturing for instance a portion of a stationary head and a right arm moving upward (see Figure 1 top-right for further examples). In addition, these patches potentially capture salient self-occlusion image patterns, e.g., dynamic occlusion. Interestingly, a similar "spacetime fragment" approach has been proposed as a theory of visual representation of actions for human perception [3]. The discovery of actemes in the learning stage is posed as a *strongly* supervised process. Specifically, hand labeled keypoint positions across space *and* time are used in a data-driven search process to discover patches that are highly clustered in the spacetime keypoint configuration space and yield discriminative, yet representative (i.e., exist in a large number of instances of the same action), patch classifiers, cf. [1]. The efficacy of this representation is demonstrated on action classification and action detection. Figure 1 provides an overview of the proposed approach.

A key benefit of the proposed representation is that it captures discriminative action-specific patches and conveys partial spatiotemporal pose information and by extension motion. Furthermore, these parts convey semantic information and the spatiotemporal location of the action. Unlike, local feature points, actemes are designed to capture the action (appearance and motion) with minimal scene context.

1.1. Contributions

In the light of previous work, the major contributions of the present paper are as follows. (i) A discriminative multiscale spatiotemporal patch model is developed, termed *actemes*, that serves as an effective mid-level representation for analyzing human actions in video. Exploiting the spatiotemporal context of actemes facilitates both action classification and detection. (ii) To realize this patch model, this work introduces a new *annotated* human action dataset[1] containing 15 actions and consisting of 2326 challenging consumer videos. The annotations consist of action class labels, 2D keypoint positions (13 in all) in each video frame and their corresponding visibilities, and camera viewpoints.

2. Acteme discovery

Human actions are often described in terms of the *relative* motion between the body parts, e.g., raise arm in front of the shoulder. Due to the large performance variation of an action, the appearance and motion cues exhibit large in-class variability. Consequently, building a part classifier directly on the initial imagery places an inordinate burden on the learning process to generalize across performance nuances. Given a set of keypoint locations, similar poses

[1]The dataset and annotations are available at the project page: http://www.scs.ryerson.ca/~kosta/actemes.html.

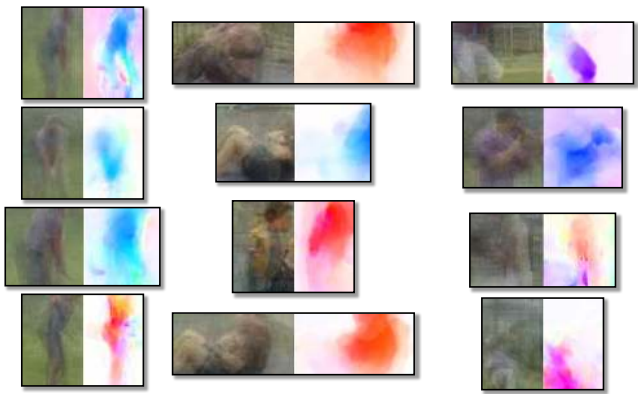

Figure 2. Example part clusters for (left-to-right) golf swing, sit up and baseball pitch, shown as the average of the color image (left) and optical flow (right) for a single frame. The consistent part alignments inside each cluster are exemplified by the sharpness in the color and motion images. This alignment leads to better classifier detection performance.

across space *and* time can be aligned which in turn reduces the variation in appearance and motion. This is exemplified in Figure 2, where the aligned action parts are relatively sharp compared to the blurred background. The introduced alignment step leads to better classifier detection performance. This approach captures the *relative* motion to the underlying pose, which transcends the traditional bags of words approach that only captures *holistic* motion.

The steps for realizing the set of actemes are as follows. The initial step consists of determining a set of patch clusters that are potentially relevant to analyzing a given set of actions (§2.1). Here, it is assumed that the keypoint locations and their visibilities are provided with the training data, see §4. Next, a set of linear classifiers are trained using the clusters as positive examples (§2.2). The final part discovery step ranks and selects a subset of actemes that are both discriminative and representative (§2.3).

2.1. Generate acteme candidates

Given the training videos, a set of random patches are selected, the seed volumes. (The training set was doubled by flipping the videos and annotation labels along the spatial horizontal axis.) For each seed volume, the optimal 2D spatial similarity transform and discrete temporal offset is estimated that aligns the keypoints spatiotemporally in each training video with the corresponding visible points within the seed volume. In addition, this fitting process is computed over a small set of temporal rescalings of the training video around unity to improve the temporal alignment. Consequently, each cluster is not only restricted to a given viewpoint and set of body poses but also a limited range of action execution speeds. Figure 3 shows an example seed volume and its three nearest patches in the training set.

Figure 3. An example seed volume is shown in the top-left panel (as frames) and the three nearest patches in the dataset are shown in the remaining panels.

The motivation behind restricting the search to a small range of temporal scales around unity is twofold. First, large temporal rescalings of a video typically introduce significant ghosting artifacts due to the relatively low sampling of the temporal dimension. Second, low and high speed patterns contain distinct natural signal statistics that may be useful for pattern matching, such as image blur in high speed patterns [11]. Thus, multiple pattern models tuned to particular speeds allows for exploiting such statistics. (Similar observations regarding image gradient statistics in natural scenes have been exploited in a recent state-of-the-art human pose estimator [31].)

Each seed volume is selected randomly from a training video at a uniformly random spatiotemporal position. Further, the spatiotemporal dimensions of the seed volume are chosen randomly from a set of four spatial dimensions, $\{100 \times 100, 100 \times 150, 150 \times 100, 100 \times 200\}$, and a set of three temporal extents, $\{5, 10, 15\}$. The random seed volume set is augmented with seed volumes centered at the temporal discontinuities of joint trajectories, since these regions of non-constant motion correspond to distinctive events that may yield discriminative patches. In total, 600 seed volumes are considered per action.

For the visible keypoints within the extents of a seed volume, P_1, the distance is computed to each video sharing the same action and camera viewpoint, as follows (cf. [1]), $D(P_1, P_2) = d_P(P_1, P_2) + \lambda_1 d_v(P_1, P_2) + \lambda_2 d_m(P_1, P_2')$, where d_P denotes the Procrustes distance between the spatial dimensions of the visible points in the seed patch, P_1, and the corresponding points in the secondary video patch, P_2, d_v measures the consistency of the visibility labels between patches, d_m measures the consistency between velocities of corresponding points in the patches after the alignment has been applied to P_2, denoted by P_2', and $\lambda_{\{1,2\}}$ are weighting factors. The visibility similarity, d_v, is defined as one minus the intersection over union of the keypoints present in both patches. The motion consistency similarity, d_m, is defined as the mean square distance between the velocities of corresponding points in the patches after the alignment has been applied.

2.2. Train acteme candidates

Similar to poselets, acteme detection is performed using a linear SVM classifier. In this work, each acteme is represented as the concatenation of the normalized and unnormalized histogram of flow (HoF), computed using dense flow [2], and the histogram of spatial gradients (HoG) [8]; alternative features are also possible, e.g., [4]. The HoF features use the same grid size as HoG, with the flow directions quantized into five bins using linear interpolation.

The negative examples are randomly selected from images that have a different action label to the positive examples. A second round of SVM training using hard false positives is also performed to improve the detection scores [1, 8]. Finally, to bring the detection scores to a comparable range, a logistic is fit over the positive and negative scores to realize a probability score.

2.3. Ranking and selecting actemes

The initial set of actemes are selected in a random fashion and thus may exhibit a wide range of capacities to *discriminate* different actions and *represent* a wide range of instances of the same action. The objective is to select a subset of the action-specific classifiers that are both discriminative and representative. This subset selection problem is combinatorial in the prohibitive large space of candidate subsets. (Note, a variety of subset selection schemes can be considered here, e.g., [1, 19].)

For each acteme candidate, its *discrimination capacity* is measured by the activation ratio, α, between the in-class and out-of-class detections among the top 1000 detections. The key idea is that distinctive actemes should fire more often on imagery of its own action category. The *representative capacity* of the i-th acteme is measured by a binary vector f_i indicating whether a detection occurs for each of the training instances. Initially, the selected acteme set, $\mathcal{M}$, contains the acteme with the highest activation ratio, α. We incrementally add the acteme candidates into $\mathcal{M}$ by alternating between the following two steps: (i) sample a training instance inversely proportional to the cumulative hit counts $\sum_{i\in\mathcal{M}} f_i$, this instance is typically underrepresented by the current $\mathcal{M}$ and (ii) pick a candidate acteme that covers the chosen training instance. If there are multiple candidates, the one with the highest α is selected and added to $\mathcal{M}$. This process is terminated when $|\mathcal{M}|$ reaches the desired maximum number.

3. Acteme inference

3.1. Acteme detection

Given an input video, each acteme is considered in a sliding volume manner, i.e., each detector is applied at all spatiotemporal positions and a set of spatiotemporal scales. To eliminate redundant part detections, non-maximum suppression is performed on the (thresholded) detector responses to yield a sparse, non-overlapping set of detections for each acteme. Each detection $\mathbf{d}_i$ is a detected spacetime volume: $\mathbf{d}_i = (\mathrm{s}_i, b_i, T_i)$, where s_i denotes the detection score, b_i is the 2D bounding box projection of the 3D volume (same at every image) and T_i is the life span of the detection. In practice, the detection scores are efficiently realized by adapting the frequency domain scheme proposed elsewhere [4].

The actemes for the same action are strongly correlated in both space and time. Exploiting the spatiotemporal context of mutual activations helps to boost the weak detections due to partial occlusion and extreme deformation. It also disambiguates similar actemes belonging to different actions, e.g., the lower body of a golf-swing vs. a baseball-swing. This is particularly important for action classification since it is desirable that each acteme only responds to instances containing its action.

Similar to [1], the detection score of each acteme is modified by looking at the detection information of other actemes in the neighboring spacetime vicinity. For each detection, a feature of length K is constructured, where the k-th entry stores the maximum detection score of the k-th acteme. This feature is concatenated with another $2 \times K$ dimensions that contain the same feature but restricted to detections occurring *earlier* and *later* in time. This feature encodes the causalities between different actemes. A linear SVM is trained using this feature on the training set and the score is converted into a probability using a logistic.

3.2. Acteme clustering via semantic compatibility

Given the detected actemes, a grouping of consistent candidates is sought that explains a consistent action hypothesis. One obvious approach is to let each candidate vote for the spatiotemporal volume of the entire action and select the best hypothesis in the voting space. This approach is inadequate because actions can undergo a large global non-linear temporal deformation. For instance, "clean and jerk" is composed of two movements, the clean and the jerk. While the two movements generally exhibit a linear temporal deformation, i.e., difference in speed, in between these movements a variable length pause is present.

Instead, this problem is addressed in a distributed fashion via figure-ground separation. First, the pairwise compatibility between the actemes is measured, where ideally the acteme detections corresponding to the same action hypothesis should be more compatible. Next, these compatibilities are used in an agglomerative clustering process to realize the spatiotemporal bounding volumes.

Given two detections $\mathbf{d}_i$ and $\mathbf{d}_j$, their semantic compatibility is computed based on the empirical distribution of keypoints, see Figure 4 left. Let $\mathcal{P}_i$ and $\mathcal{P}_j$ be the distribu-

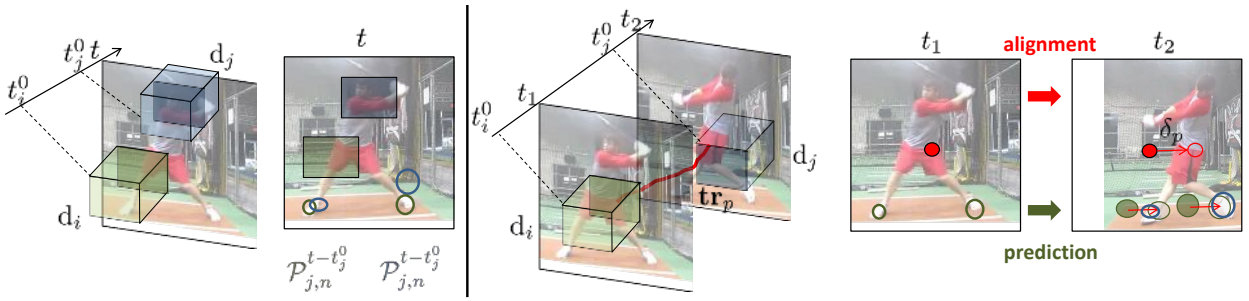

Figure 4. Semantic compatibility computation between two actemes (left) overlapping in time and (right) not overlapping in time. Left: acteme $\mathbf{d}_i$ (green) and $\mathbf{d}_j$ (blue) overlap at a frame t. The consistency is checked by comparing the empirical distribution of the keypoints of $\mathbf{d}_i$ (green circle) and $\mathbf{d}_j$ (blue circle). Right: trajectory $\mathbf{tr}_p$ goes through two actemes (green) not overlapping in time. The keypoint distribution of $\mathbf{d}_i$ at t_1 is propagated into the future frame t_2 (shaded green) and translated by the alignment δ_p induced by the trajectory.

tions of all N annotated keypoints in the training set of the actemes. We define the the semantic compatibility score as:

$$c(\mathbf{d}_i, \mathbf{d}_j, t, \delta) = \frac{1}{N} \sum_{n=1}^{N} D_{KLD}(\mathcal{P}_{i,n}^{t-t_i^0}, \mathcal{P}_{j,n}^{t-t_j^0} + \delta), \quad (1)$$

where D_{KLD} is the symmetric KL-distance, i.e., the sum of two KL-divergences where the distribution inputs are flipped, t_i^0, t_j^0 denote the starting frame of two detections and δ measures the relative displacement where we measure the KL-divergences.

There are two cases to consider for measuring the compatibilities: (i) two detections overlap in time and (ii) two detections do not overlap in time. For two acteme detections overlapping in time, the semantic compatibility score is computed by taking the average over their common frames:

$$C_s(\mathbf{d}_i, \mathbf{d}_j) = \frac{1}{|T_i \cap T_j|} \sum_{t \in T_i \cap T_j} c(\mathbf{d}_i, \mathbf{d}_j, t, 0). \quad (2)$$

For two acteme detections not overlapping in time, the underlying actemes could capture different stages in time. To measure the consistency between them requires (i) prediction of the keypoint configuration in the future and (ii) alignment to deal with camera movement (image coordinate change) or human movement. Prediction is accomplished in a data-driven manner by propagating the ensemble keypoint distribution of the training instances into the future. For the alignment, the point trajectory is used as an anchor point. Each trajectory $\mathbf{tr}_p$ is a series of spacetime points, and $\mathbf{tr}_p^t$ is its position at time t. If $\mathbf{tr}_p$ goes through the volume of $\mathbf{d}_i$ and $\mathbf{d}_j$, we can translate the keypoints using the trajectory displacement, see Figure 4 right. The point trajectories are obtained by linking the optical flow fields in time.

Specifically, let t_1, t_2 be the closest frames between two detections, such that $\mathbf{tr}_p^{t_1} \in b_i$ and $\mathbf{tr}_p^{t_2} \in b_j$. We use $\delta_p = \mathbf{tr}_p^{t_1} - \mathbf{tr}_p^{t_2}$ to denote the translation of the trajectory. The induced semantic compatibility score is: $C_t(\mathbf{d}_i, \mathbf{d}_j, \mathbf{tr}_p) = c(\mathbf{d}_i, \mathbf{d}_j, t_2 - t_i^0, \delta_p) + c(\mathbf{d}_i, \mathbf{d}_j, t_1 - t_j^0, \delta_p)$. We calculate the semantic compatibility score, $C_s(\mathbf{d}_i, \mathbf{d}_j)$, as the average of C_t over all common trajectories, weighted by their length, since longer trajectories are less likely to yield erroneous correspondences.

If two actemes are far apart in time, the prediction tends to be unreliable and the compatibility is set to zero. As the acteme detections are dense, the transitivity of compatibility among them links distant frames.

Given the computed semantic compatibility score, C_s, clustering is achieved in an agglomerative manner, cf. [1]. The first step selects the detected acteme with the highest score. When the average linkage score falls below a parameter τ, a new cluster is formed. The clustering process is terminated when the detection score, s, falls below the threshold, ρ.

For each acteme cluster, the constituent actemes predict the human bounding volume within its life span. Next, the predicted volume for each cluster is merged to form an action detection. For the actemes that are closest to the beginning (end) of the action, a predictor for the start (end) time is trained for the whole action and is used to refine the detection at test time.

4. Crowdsourced mocap

In object recognition, much effort has been placed on the collection of richly annotated image databases in natural settings. Such metadata have proven extremely useful in supervised training of object models (e.g., [8, 1]). In contrast, traditional action-related datasets (e.g., KTH [27]) have generally been comprised of staged videos in limited settings, such as limited camera viewpoint, occlusion and background clutter. Recently, there have been efforts to collect large datasets from consumer sources to address these issues [17, 23, 16]. While this is a positive trend, these datasets, as their predecessors, generally lack training metadata, such as bounding boxes and keypoint annotations.

Critical to realizing a set of actemes are the locations and visibility labels for a set of keypoints throughout the

action performance. (Note that this metadata is exclusively used in the *learning* process.) In this work, VATIC [28], an open source semi-automated video annotation tool designed for labeling multiple "object" tracks in a video, was adapted and deployed on Amazon Mechanical Turk (AMT) to annotate the videos in the introduced dataset. Three independent annotations were collected for each video and combined via outlier rejection and averaging. Workers were tasked with providing the labels for the following 13 keypoints: head, left-right shoulder, left-right elbow, left-right wrist, left-right hip, left-right knee and left-right ankle.

The videos in the introduced dataset were obtained from various online video repositories, such as YouTube (www.youtube.com). The dataset consists of the following 15 action classes: baseball pitch, baseball swing, bench press, bowling, clean and jerk, golf swing, jump rope, jumping jacks, pull up, push up, sit up, squat, strum guitar, tennis forehand and tennis serve. Challenging aspects of the dataset include large variation in intra-class actor appearance, action execution rate, viewpoint, spatiotemporal resolution and complicated natural backdrops. The videos were temporally segmented manually such that they contained a single instance of each action.

5. Empirical evaluation

The training/testing data is split 50/50. Clips in the training (includes spatial horizontal flipping) and testing sets are selected such that clips from the same video sequence do not appear in *both* sets. During training, the configuration distance is computed with $\lambda_1 = \lambda_2 = 0.1$. For each action category, 300 actemes are selected ($K = 4500$ in total). During testing, acteme detection is run at eight different spatial scales, four per octave and three temporal scales, i.e., $\{0.9, 1, 1.1\}$. Non-maximal suppression is done for acteme detections surpassing a 50% spatiotemporal overlap ratio. The agglomerative clustering is stopped when τ falls below 0.2. The detection threshold, ρ, is chosen such that 90% of the true positives are above the threshold. To facilitate the optical flow computation (about 1fps), the videos are resized to a maximum dimension of 480 pixels.

5.1. Action classification

The effectiveness of the proposed model is first shown for action classification. Several standard local features combined in a bag of words model are used as baselines. The protocol from a recent action recognition evaluation of feature descriptors is followed [30]. Each video is represented by the histogram of frequencies for each word. In training and testing, features are assigned to their closest vocabulary word based on the Euclidean distance. A one-vs-all SVM classifier is trained for each category. At test time, the action label corresponding to the detector with the maximum score is returned as the label. Several kernels are used with the intersection kernel (IK) providing the best performance overall. All presented results herein are based on IK.

		inside bbox	outside bbox	global
STIP [30]	HoG	62.5%	33.2%	61.3%
	HoF	83.9%	54.2%	82.5%
	HoG + HoF	84.9%	50.8%	82.9%
Dense [30]	HoG	62.6%	47.4%	54.5%
	HoF	83.8%	64.2%	77.4%
	HoG + HoF	83.7%	55.9%	73.4%
Cuboid [5]		64.5%	48.2%	67.5%
HoG3D [14]		85.3%	52.3%	84.5%
MIP [15]		76.1%	61.5%	78.0%
Action Bank [25]		90.6%	-	83.9%
Actemes		88.0%	-	79.4%

Table 1. Comparison of mean average precision across all categories for various extant image cues. Action Bank is not directly comparable since it leverages additional training data outside of the introduced dataset to realize their representational substrate.

Comparison is made with the spatial histogram of gradients (HoG), histogram of optical flow (HoF) and their concatenation (HoG+HoF) [30]. These features are computed both densely and on extracted spacetime interest points. We also compare with spacetime cuboids [5] on interest points, dense spacetime histogram of gradients (HoG3D) [14], and dense local motion interchange pattern [15]. A bag of words model is run under three different scenarios: (i) inside the action bounding volume, (ii) outside bounding volume of action and (iii) the whole video clip. As an additional baseline, we compare against Action Bank [25] which represents actions on the basis of holistic spacetime templates.

If the detection window is provided, as considered in the VOC Action Classification challenge [19], actemes whose predicted spatiotemporal action bounds overlap sufficiently with the input action boundary are retained for further processing (cf. [19]); the overlap ratio is defined by the intersection over union of the input and predicted action volume boundaries. We use 75% as the overlap ratio threshold. The introduction of the overlap criteria enforces a degree of spatiotemporal translation invariance and spatiotemporal organization among the detected actemes within the analysis volume and suppresses spurious detections. The ensemble of sparse detections provide the intermediate substrate for further processing.

To classify an action, a fixed-length activation vector representation is used [19], where the dimensionality of the vector corresponds to the number of actemes. The maximum response for each acteme is used as the entry in the corresponding component of the activation vector, i.e., max-pooling. For each action, a separate linear support vector machine (SVM) classifier is trained on top of the vector of acteme responses.

Results are compared in Table 1 and the confusion matrix for the acteme-based classification (given the bounding volume) is given in Figure 5. The proposed approach

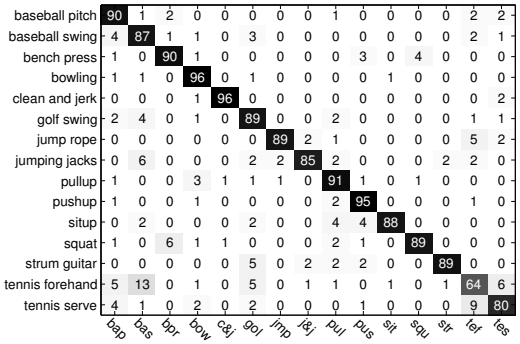

Figure 5. Confusion matrix for acteme-based classification given the bounding volume of the action.

achieves a better result than low-level features when the bounding box is provided, i.e., scene context is limited and thus performance is not conflated with non-action information. Furthermore, the largest gain of the proposed approach over interest points is found in separating actions performed in similar scenes (e.g., bench press vs. squat). This is indicative of the proposed representation being sensitive (by design) to the pose. When the bounding boxes are not provided, the false positives of the actemes on the background contaminate the activation signal and thus reduce performance. This can be remedied by clustering the top detection (as done in §5.2).

5.2. Action detection

In addition to classification, the proposed model is evaluated on action detection. Specifically, given an input video, the goal is to determine *where* and *when* one of the 15 actions is performed. This is captured by a 3D spacetime bounding volume.

In evaluation, detection is measured using precision-recall. For each action category, the corresponding actemes are run on the input video. Detected actemes are clustered using the semantic compatibility, and non-maximal suppression is performed using a 50% overlap ratio. The score of each detection s_i is computed as the sum of the clustered acteme detection scores. During testing, all the detection hypotheses are ranked across the entire dataset and the precision-recall curve for each category is computed, see Figure 6. A detection hypothesis that overlaps with the ground truth with a ratio greater than 50% is treated as a true positive.

Comparison is made with Hough Forest [10]; the model is retrained on the introduced dataset using the provided code. Each sequence is subsampled by the first, middle and last frame. Then 15 Hough trees are trained, five at a time. During testing, the top detections from these three frames are linked based on aspect ratios and trajectories.

The proposed detector achieves an average precision of 0.45, compared to 0.26 achieved by the baseline. This demonstrates the discrimination power of the proposed representation, especially with scene clutter. Figure 6 shows several example detections.

6. Discussion and summary

In this paper, a novel approach for analyzing human actions in video was presented. The key concept introduced is modeling human action as a composition of discriminative volumetric patches, termed actemes. Each patch, found through a data-driven discovery process, implicitly captures the intricate pose, dynamics, self-occlusion and spatial appearance of a subset of body parts in a spatiotemporal neighborhood. The discovery of these parts relies on strong supervision, e.g., the position of keypoints across space-time. While similar levels of supervision are common in the context of analyzing static imagery (e.g., [1, 31]), such supervision has not previously been fully exploited in the context of human actions in non-scripted, unconstrained video settings. This approach allows to spatiotemporally localize an action in cluttered scenes by a bounding volume. Possible future work includes extending the framework to modeling person-object interactions, and segmenting out actions at the pixel level for a finer-grained analysis.

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[3] T. Davies and D. Hoffman. Facial attention and spacetime fragments. *Axiomathes*, 13(3-4):303–327, 2003.

[4] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *PAMI*, 35(3), 2013.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

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

[7] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[9] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011.

[10] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 33(11):2188–2202, Nov. 2011.

[11] W. Geisler. Motion streaks provide a spatial code for motion direction. *Nature*, 400(6739):65–69, 1999.

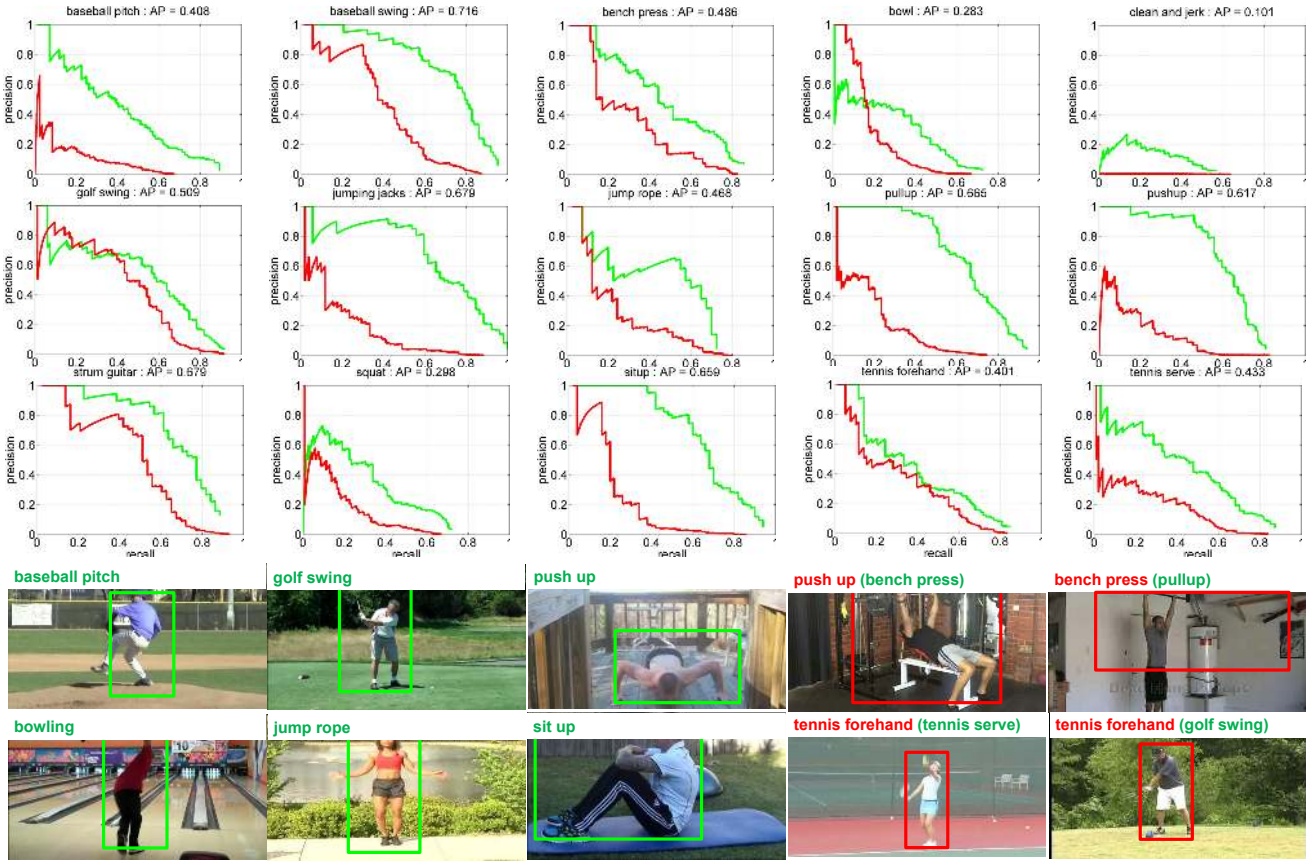

Figure 6. Detection results. Top three rows show the precision-recall curve for our detection approach (green) and Hough forest baseline (red) for each action. Bottom two rows show example true (green) and false (red) positives.

[12] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

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

[14] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[15] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012.

[16] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *PAMI*, 34(3):615–621, 2012.

[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] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[20] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors. *Visual Analysis of Humans - Looking at People*. Springer, 2011.

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

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

[23] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.*, 24(5):971–981, 2013.

[24] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

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

[26] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? In *CVPR*, 2008.

[27] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, pages III: 32–36, 2004.

[28] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2012.

[29] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

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

[31] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.