Poselet Key-framing: A Model for Human Activity Recognition

Michalis Raptis and Leonid Sigal
Disney Research, Pittsburgh
{mraptis,lsigal}@disneyresearch.com

Abstract

In this paper, we develop a new model for recognizing human actions. An action is modeled as a very sparse sequence of temporally local discriminative keyframes – collections of partial key-poses of the actor(s), depicting key states in the action sequence. We cast the learning of keyframes in a max-margin discriminative framework, where we treat keyframes as latent variables. This allows us to (jointly) learn a set of most discriminative keyframes while also learning the local temporal context between them. Keyframes are encoded using a spatially-localizable poselet-like representation with HoG and BoW components learned from weak annotations; we rely on structured SVM formulation to align our components and mine for hard negatives to boost localization performance. This results in a model that supports spatio-temporal localization and is insensitive to dropped frames or partial observations. We show classification performance that is competitive with the state of the art on the benchmark UT-Interaction dataset and illustrate that our model outperforms prior methods in an on-line streaming setting.

1. Introduction

It is compelling to think of an action, or interaction with another person, as a sequence of keyframes – key-poses of the actor(s), depicting key states in the action sequence. This representation is compact and sparse, which is desirable computationally and for robustness, yet is rich and descriptive. The sparsity and compactness come from the fact that keyframes are, by definition, temporally *very* local, in our case, each spanning just two frames (using the second frame to compute optical flow).

It is worth noting that this use of local temporal information is in sharp contrast to most research in video-based action recognition where often long temporal trajectories [24] or features computed on much larger temporal scale (20 or 100 frame segments [32]) are deemed necessary. Using a sparse local keyframe representation, however, does have certain benefits. First, it allows our model to focus on the most distinct parts of the action and disregard frames that are not discriminative or relevant. Second, it translates to robustness to variation in action duration or dropped frames because these changes minimally affect our representation. Further, in perception, it has long been shown that certain discriminant static images of humans engaged in activity can convey dynamic information (an effect known as *implied motion* [15][1]). These studies, along with the success of keyframe summaries as means of motion illustration [1] and/or synthesis in computer graphics, motivate us to consider local keyframes as sufficient for our task. However, discovering such keyframe representations is challenging because, intuitively, it requires having accurate pose information for the entire video [30], which is both difficult and computationally expensive.

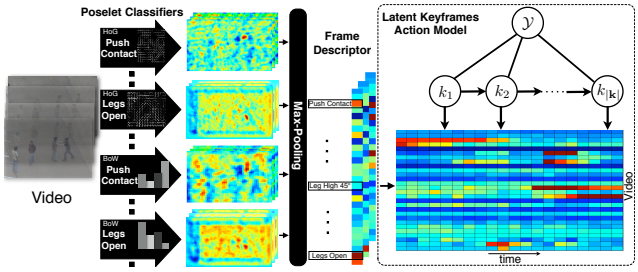


Figure 1. **Model framework:** Image frames are encoded by poselet activation max-pooled over the spatial extent of each frame. An action is modeled by a set of latent keyframes discriminatively selected using a max-margin learning framework.

Motivated by the success of *poselets* in human detection [3] and pose estimation [35] we posit that representing a keyframe as learned collection of *poselets* has a number of significant benefits over the more holistic person/pose-based representation [34]. The key benefit of poselets is that they can capture discriminative action parts that carry partial pose (and, in our case, motion) information; this is extremely useful in complex environments where there is clutter or the actor experiences severe occlusions. Moreover, poselets also allow for a semantic, spatially localizable, mid-level representation of the keyframe.

[1] Further, it has been shown that the same parts of the brain that engage in processing and analysis of motion are engaged in processing implied motion in static images [15].

For video-based action recognition, which is the goal of this work, one keyframe may not be sufficient to adequately characterize the action. How many distinct (key)frames are needed to characterize a human activity [29]? For instance, a handshake between two people can be summarized using 3 distinctive keyframes: (i) the two persons approach each other, (ii) they extend their hands towards one another when near, and (iii) they touch and shake their hands. While it may be sensible to specify the number of keyframes to use in a representation; specifying the location of such keyframes for many actions would be too tedious and, frankly, prone to errors. Humans are notoriously bad at the task, and semantic meaningfulness may not translate well to discriminability.

Contributions: We cast the learning in a max-margin discriminative framework where we treat keyframes as latent variables. This allows us to (jointly) learn a set of the most discriminative keyframes while also learning the local temporal context between them. Our model has appealing properties. First, it allows temporal localization of actions and action parts by modeling actions as sequences of keyframes. Second, it is tolerant to variations in action duration, as our model only assumes partial ordering between the keyframes. Third, it implicitly allows spatial localization of actions within keyframes by representing the keyframes with *poselets*. Fourth, our formulation is amenable to on-line inference for early detection [22]. Finally, our model generates semantic interpretations (summarizations) of actions by modeling them as *action stories* – contextual temporal orderings of discriminant partial poses.

1.1. Related Work

The problem of action recognition has been studied extensively in the literature. Given the significant literature in the area, we focus only on the most relevant works.

Sequential models: Hidden semi-Markov Models (HSMM) [10] , CRFs [31], and finite-state-machines [13] have been used to model the temporal evolution of human activities. Recently, Tang *et al*. [32] propose a conditional variant of HSMM incorporating the max-margin framework in the training phase. These works model entire video sequence both its progression and the temporal duration of each phase. As consequence, during training those models need to also encode irrelevant to the action events, making the learning procedure inherently challenging. Our model is more flexible, selecting and modeling only a very sparse, discriminative subset of the sequence.

Static image activity recognition: Most approaches in still image action recognition assume a single actor and rely on either explicit [8, 33] or implicit pose [36] information and, often, bag-of-words (BoW) terms for background context [8, 36]. For example, in [33] actions are represented by histograms of pose primitives. Delaitre *et al*. [8] uses a more traditional part-based pose model instead. These methods, yet, implicitly inherit the problems of traditional pose estimation. Yang *et al*. [36] and Maji *et al*. [21] side-step these problems by proposing to use poselets as a mid-level representation for pose; in [21] poselet activation vectors are used directly for recognition, whereas in [36] intermediate latent 3D pose is constructed to bootstrap performance. We leverage a poselet-like mid-level representation for the frames, but focus on a video scenario where multiple discriminative frames must be selected to encode the action.

Key volumes: Recent video-based action recognition methods [12, 28] observed that limiting the bag of spatio-temporal interest point representation [9, 17] in a temporal segment boosts recognition performance. Niebles *et al*. [23] combines the BoW representation of the entire video (global term) with sub-volumes that capture the temporal composition of the action. However, the proposed model lacks the ability to spatially localize action parts. Moreover, the model of [23], similar to [24], relies on global terms, assuming that rough temporal segmentation is given. In contrast, we propose an extremely local action model geared towards analyzing longer image sequences. Brendel and Todorovic [4] introduce a generative model that describes the spatio-temporal structure of the action as a weighted directed graph defined on a spatio-temporal over-segmentation of the video. Chen and Grauman [6] propose an irregular sub-graph model in which local temporal topological changes are allowed.

While key volume methods focus on discriminative portion of the video, their volumetric nature is susceptible to variabilities present in the temporal execution of the action. In contrast, our method decouples action representation from its exact temporal execution, focusing only on temporally local keyframes that are less variable.

Keyframes: A number of approaches have proposed the use of keyframes as a representation. Carlson *et al*. [5] use a single, manually selected, keyframe and shape matching to classify tennis strokes; [39] rank all frames based on holistic information theoretic measure to select the top 25% for classification using voting; [20] rely on spatio-temporal localization as pre-processing and use AdaBoost to select keyframes (making up from 13% to 20% of the sequence length). Unlike [5], we automatically select keyframes and return, not require as [20], spatial-temporal localization; our keyframe representation is also more compact utilizing fewer (up to 4, or 4% of the sequence) keyframes.

Vahdat *et al*. [34] propose a max-margin framework for modeling interactions as sequences of key-poses performed by a pair of actors. To model interactions, the approach requires complete tracks of both actors across the entire sequence. In contrast, we rely on a collection of poselets to characterize frames and hence can better deal with partial occlusions, and we are not limited to interaction scenarios.

In addition, since keyframes in [34] are built as associations to full-body exemplars, the space of possible keyframes is linear in those exemplars; with our distributed poselet representation the space of possible keyframes is exponential in the number of poselets, giving us a more expressive model.

Attributes: This class of methods incorporate rich human knowledge to create mid-level representations that capture intrinsic properties of atomic movements (*e.g.*, "moving lower body", "still torso"). Liu *et al.* [19] generate a representation of a given video based on the detected active attributes. Kong *et al.* [14] construct similar but human-centric–representations, assuming ideal person tracking in each video. The attribute-based models do not capture the spatio-temporal structure of the action, however, incorporate expert knowledge. Our model tries to close this gap by building a localizable mid-level representation.

2. The Model

A graphical representation of our model is illustrated in Fig. 1. This model has the ability to localize temporally and spatially the discriminative action components. It performs action classification and generates detailed spatial and temporal reports for each component. The temporal context of the action, such as the ordering of the action's components and their temporal correlations are explicitly modeled. Moreover, the model implicitly performs spatial localization of the image regions that are part of the action.

2.1. Model Formulation

Given a set of video sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathcal{X}$ and their associated annotations $\{y_1, \ldots, y_n\}$, with $y_i \in \{-1, 1\}$, our purpose is to learn a mapping $f : \mathcal{X} \rightarrow \{-1, 1\}$. This mapping function will also enable the automatic temporal annotation of unseen video sequences. Our input variables are sequence of images $\mathbf{x}_i = \{x_i^1, \ldots, x_i^{T_i}\}$, where T_i is the length of the video. Our output variable consists of a "global" label y indicating whether a particular action occurs inside the video. Additionally, we introduce auxiliary *latent* variables $\mathbf{k} \in \mathcal{K}$, where $\mathcal{K} = \{k_i \in \mathbb{Z}_+ : k_i < k_{i+1}\}$. Those *latent* variables specify the subset of frames that our model considers. Hence, our hypothesis y^* is: $y^* = \text{sign}(f(\mathbf{x}; \mathbf{w})) = \text{sign}(\max_{\mathbf{k} \in \mathcal{K}} F(\mathbf{x}, \mathbf{k}; \mathbf{w}))$. Our scoring function is written as the sum of unary and pairwise terms:

$$F(\mathbf{x}, \mathbf{k}; \mathbf{w}) = \sum_{i=1}^{|\mathbf{k}|} \langle w_i, \phi(x^{k_i}) \rangle + \sum_{i=1}^{|\mathbf{k}|-1} \langle w_{(i,i+1)}, \psi(x^{k_i}, x^{k_{i+1}}) \rangle + b \,. \quad (1)$$

Both unary and pairwise terms extract information from the video sequence sparsely. Only the frames indicated by the latent variables $\mathbf{k}$ contribute to the scoring of the video.

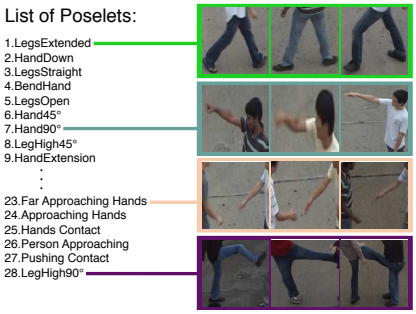


Figure 2. *Examples of annotated poselets.*

3. Frame Descriptor

We formulate our frame descriptor based on the notion of *poselets* [3] – localizable discriminant parts of the body or, in our case, action. Based on a few weak annotations on a sparse set of frames (Fig. 2), we build two types of poselets using Histograms of Oriented Gradients (HoG) templates [7] and BoW features. The original poselets [3], as well as those used in static action recognition [21], rely purely on HoG descriptors. HoG provides an effective, localizable representation based on the (global) structure of the edge information. This comes at a loss of fine (local) appearance details that get washed away due to the rigid structure of the template. Further, HoG templates do not model motion, which is an important cue for action recognition.

To address the aforementioned limitations, we augment HoG-based poselets with BoW poselets (both trained from the same annotations). BoW features, quantized dense descriptors (SIFT, Histogram of Optical Flow (HoF), and Motion Boundaries (HoMB)) [24], are most often used as holistic descriptors of images or spatio-temporal volumes. In contrast, we use BoW features to characterize poselets, allowing us to have a spatially localizable representation (similar to work in object detection [2], but with BoW computed over appearance and motion). Our poselet representation gives us a flexible framework where we are able to define semantic poselets that are generic (spanning multiple activities) and are also action specific.

For describing a frame of a video, we collect the highest scores from each poselet classifier (both HoG and BoW based) and form a poselet activation vector [21]. Finding the bounding box with the highest score in the case of BoW representation can be done very efficiently using branch-and-bound techniques [16]. The HoG-based poselets are evaluated in a scanning window fashion and the highest response is stored. Our final feature representation $\phi(x^{t_i}) \in \mathbb{R}^{2M}$ is formed by concatenating the max-pooled [18, 27] M poselet activation scores. Consequently, each component of a unary term $w_{i_j} \phi(x^{k_i})_j$ (Eq. 1) scores the compatibility of the activation score of j-th poselets with the i-th keyframe of the action model.

Learning of poselets: Learning the poselet classifiers is a challenging task, because in our framework, unlike in [3], we do not assume good alignment of the annotations based on the joint locations of the actors or provide an explicit negative set. Instead, we rely on structured output learning, proposed by Blaschko and Lampert [2] for localizing objects in an image, to mine for hard negatives and align our weakly annotated poselet templates. This ensures that learned poselets are effective for spatial localization.

Given a set of images $\{I_1, \ldots, I_n\}$ and their annotations $\{\hat{y}_1, \ldots, \hat{y}_n\} \subset \mathcal{Y}$ for a specific type of poselet, we wish to find a mapping function that can localize the same poselet in an unseen image. The structured output is the space of bounding boxes or no bounding box in an image:

$$\begin{aligned} &\underset{\beta,\xi}{\text{minimize}} \ \frac{1}{2}\|\beta\|^2 + C^{'} \sum_{i=1}^{n} \xi_i \qquad (2) \\ &\text{s.t } \langle \beta, \theta(\mathbf{I}_i, \hat{y}_i)\rangle - \langle \beta, \theta(I_i, \hat{y})\rangle \geq \Delta(\hat{y}_i, \hat{y}) - \xi_i, \ \forall \hat{y} \in \hat{\mathcal{Y}} \\ &\quad \xi_i \geq 0, \ \ \forall i \end{aligned}$$

where $\Delta(\hat{y}_i, \hat{y})$ is, for the positive images, a loss function that encodes the amount of overlap the predicted bounding box $\hat{y}$ has with the ground truth $\hat{y}_i$. For the negative images, $\Delta(\hat{y}, \hat{y}) = 1$, if the prediction indicates a poselet is present. Moreover, we set $\theta(\mathbf{I}_i, \hat{y}_i) = \mathbf{0}$ for the negative images. As a result, a poselet is assumed present if its detection score $\langle \beta, \theta(I_i, \hat{y})\rangle$ is above zero. For the case of BoW based poselets, the feature function $\theta(I_i, \hat{y})$ represents the concatenation of three histograms formed by the quantized dense descriptors that are contained inside the bounding box $\hat{y}$. For the case of the HoG template, the feature $\theta(I_i, \hat{y})$ corresponds to the vectorized HoG template starting from the upper left corner of the bounding box.

3.1. Pairwise Correlations

To model the temporal structure of the human activity, we encode the pairwise correlations between poselets in consecutive keyframes. We capture presence/absence of a poselet in one keyframe and the simultaneous presence/absence of another poselet in the next keyframe. Therefore, we first augment each frame descriptor ϕ (Sect. 3), decoupling the detection scores that indicate presence (positive) and absence (negative). A non-negative sparse descriptor is created:

$$\begin{aligned} \widehat{\phi(x^{t_i})} = &[\phi(x^{t_i})_1 \mathbb{1}_{[\phi(x^{t_i})_1 > 0]}, -\phi(x^{t_i})_1 \mathbb{1}_{[\phi(x^{t_i})_1 \leq 0]}, \ldots, \\ &\phi(x^{t_i})_{2M} \mathbb{1}_{[\phi(x^{t_i})_{2M} > 0]}, -\phi(x^{t_i})_{2M} \mathbb{1}_{[\phi(x^{t_i})_{2M} \leq 0]}] \end{aligned}$$

where $\mathbb{1}_{[.]}$ is the indicator function. By computing and then vectorizing the outer product of those augmented descriptors of two frames, we get our pairwise descriptor: $\psi(x^{t_i}, x^{t_j}) = vec(\widehat{\phi(x^{t_i})}\widehat{\phi(x^{t_j})}^T) \in \mathbb{R}^{4M^2}$. Based on this descriptor, we can quantify the fit to our action model of the presence of the j-th poselet in the i-th keyframe and the absence of the k-th poselet in the next keyframe by the pairwise component: $w_{(i,i+1)_{[2j-1,2k]}} \psi(x^{k_i}, x^{k_{i+1}})_{[2j-1,2k]}$.

Having our action model, we now describe two main algorithmic steps: i) selection of the optimal keyframes, and ii) learning the model parameters from training data.

4. Model Inference

Given a video sequence $\mathbf{x}_i$ and activity model $\mathbf{w}$ (Sect. 2.1) the classification process involves the maximization of our scoring function over the latent variables $\mathbf{k}$. Our model has a temporal ordering constraint on the set of keyframes selected from the image sequence, stating that the latent variables are a strictly increasing sequence of positive integers $k_1 < k_2 < \ldots < k_{|k|}$ (Sect. 2.1). These constraints enable the optimal keyframe selection to be computed by dynamic programming (DP).

Let $D(n, m)$ be the optimal value of the scoring function Eq. 1 in the case that the last of the n keyframes to be selected is the m-th frame of the image sequence. Then, based on the monotonicity constraints, we can define the following DP-equations: $D(1, m) = \langle w_1, \phi(x^m)\rangle$,

$$\begin{aligned} D(n, m) = &\max_{n-1 \leq p < m} \{D(n-1, p) + \langle w_{(n-1,n)}, \psi(x^p, x^m)\rangle\} \\ &+ \langle w_n, \phi(x^m)\rangle. \end{aligned}$$

The optimal solution is given by Eq. 3, and the indices for the keyframe are retrieved with backtracking.

$$F(\mathbf{x}_i, \mathbf{k}^*; \mathbf{w}) = \max_{|\mathbf{k}| \leq k_{|\mathbf{k}|} \leq T_i} D(|\mathbf{k}|, k_{|\mathbf{k}|}). \qquad (3)$$

The computation cost of the dynamic programming is $O(|\mathbf{k}|T_i + T_i^2) = O(T_i^2)$. The dominant computational cost term is the evaluation of the T_i^2 pairwise potentials. However, the evaluation of all the pairwise potentials is not required during testing. The maximum temporal distance between two successive keyframes can be computed during training, and a loose upper bound $\tau < T_i$ can be defined for each learned action model, leading to a more efficient inference algorithm $O(\tau T_i)$. Moreover, note that the dynamic programming inference algorithm can be used efficiently in a streaming scenario, see Sect. 6.2.

5. Learning

Our scoring function F for a video $\mathbf{x}_i$ is the inner product $\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{k})\rangle$ of the high-level features $\Psi(\mathbf{x}_i, \mathbf{k})$ and the parameter vector $\mathbf{w}$, where

$$\begin{aligned} \mathbf{w} &= [w_1, \ldots, w_{|\mathbf{k}|}, w_{(1,2)}, \ldots, w_{(|\mathbf{k}|-1,|\mathbf{k}|)}, b], \\ \Psi(\mathbf{x}_i, \mathbf{k}) &= [\phi(x^{k_1}), \ldots, \phi(x^{k_{|\mathbf{k}|}}), \psi(x^{k_1}, x^{k_2}), \ldots, \\ &\quad \psi(x^{k_{|\mathbf{k}|-1}}, x^{k_{|\mathbf{k}|}}), 1]. \end{aligned} \qquad (4)$$

Our goal is to compute the model parameters $\mathbf{w}$ that minimize the regularized risk defined here:

$$\begin{aligned} &\underset{\mathbf{w},\xi}{\text{minimize}}\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t}\ \ y_i\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{k}_i^*(\mathbf{w}))\rangle \geq 1-\xi_i,\ \ \forall i, \\ &\quad \xi_i \geq 0,\ \ \forall i. \end{aligned} \tag{5}$$

Optimizing Eq. 5 and finding the global minimum is a hard problem, because it is not a convex objective function due to the dependency of the latent variables $\mathbf{k}_i^*(\mathbf{w})$ on $\mathbf{w}$. However, we can find a local minimum by employing an alternation optimization, similar to Felzenszwalb *et al.* [11]. More specifically, we adapt the following approach:

1. Given $\mathbf{w}$: For each video, the latent variables are updated, $\mathbf{k}_i^* = \text{argmax}_{\mathbf{k}}\, F(\mathbf{x}_i, \mathbf{k}; \mathbf{w})$ (Sect. 4). For the positive samples, we fix their latent variables to $\mathbf{k}_i^*$. For the negative samples, we update their growing list of possible latent variables.

2. Given the updated set of the latent variables $\mathbf{k}_i^*$: Optimize Eq. 5 over $\mathbf{w}$.

For initialization, we randomly initialize the unary weights w_i and set the pairwise weights $w_{i,j}$ to zero. Due to the non-convexity of the problem, our learning algorithm can get trapped in poor local optima. Therefore, we repeat the training procedure a number of times and select as the final solution the model parameters that lead to the smallest objective function. In practice, we repeated the training phase with 4 different random initializations.

6. Experimental Results

We validate the proposed model on the UT-Interaction Set #1 benchmark dataset [26] using the segmented and the continuous execution versions of the dataset for the classification and temporal detection experiments, respectively. This dataset contains 10 video sequences (720×480, 30fps) that are continuous executions of 6 classes of person-person interactions: "handshaking", "hugging", "kicking", *etc.* These high-level complex actions involve combinations of discriminative atomic movements as well as interactions between humans. Therefore, holistic BoW representations of the video perform poorly [26] relative to high-level description-based methods [4, 14, 34] (Table 1). Our model[2] is evaluated based on the 10-fold leave-one-out cross-validation previously proposed in [26].

We annotated the UT-Interaction dataset with 598 bounding boxes of poselets. Fig. 2 shows examples of the annotations provided for training our poselet models (Sect. 3). Our annotations contain $M = 28$ types of generic and action specific poselets, *e.g.*, "legs extended", "hand extended 90°", "hug", "handshake contact", *etc.* Generic poselets like a "legs extended" poselet may not have one-to-one correspondences with a specific action. However, their co-occurrence with other poselets captures salient information for action discrimination. During training phases, the number of positive samples for each poselet type varies between 10 and 65. This limited amount of samples, in combination with weakly aligned annotations (Sect. 3), makes mining of hard negatives [11] a crucial step of the training procedure.

Experimental Details: The penalty parameter C of the latent SVM objective (Eq. 5) is selected with 3-fold cross-validation in each of the training sets. In contrast, the penalty parameter C' of the structure SVM objectives (Eq. 4) is set to 10 for both the HoG template and the BoW for the poselet classifiers. The vocabulary size for the dense SIFT, HoF and HoMB descriptors is set to 500.

6.1. Action Recognition

For each action class, we train our model using the same number of keyframes $|\mathbf{k}|$. A multi-class linear SVM is used to combine the scores obtained from the 6 action models to compensate for the different bias of each model. Using only 4 keyframes, our model achieves an average recognition rate of 93.3% (Fig. 3 (a)). Table 2 summarizes the results of different variations of our model along with two baseline algorithms: we use RBF-χ^2 SVM classifiers a) on a volumetric max-pooled poselet feature representation of the entire video, following [27] and b) on a BoW representation of the video based on our quantized dense descriptors. The volumetric max-pooled poselet representation directly builds on our high-level features, except that it lacks the ability to capture fine temporal information about the action. We note that our model significantly outperforms the latter approach by more than 15% in average accuracy. Even using only a single keyframe (Fig. 3 (b)), a 3% increase in accuracy is observed. Moreover, our results indicate a significant boost in performance using the combination of HoG template and BoW based poselets, proving their complimentary detection performance. Fig. 3 (b) shows the performance of our framework using the full model and only unary terms while varying the number of latent variables $|\mathbf{k}|$. We attribute the slight drop in performance using 5 keyframes to over-fitting during the training phase as the number of parameters in our model grows linearly with the numbers of keyframes.

Additionally, comparisons with other recent approaches are shown in Table 1. Our approach outperforms most other approaches and its performance is comparable with the state-of-the-art method [38]. In contrast with Vahdat *et al.* [34], our approach does not require accurate human tracking. Moreover, our model is not constructed explic-

[2]The poselet learning framework follows the same experimental setup and never sees any of the test data.

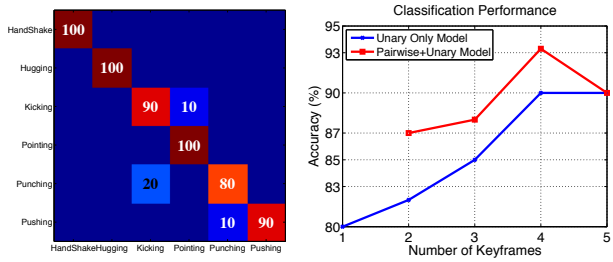
itly to model person-to-person interactions. The performance of our method is lower than the bag of higher order co-occurrence of spatio-temporal features proposed by [38]. This can be attributed to the use of a linear classification scheme compared to non-linear kernels. Although, we note that holistic representations of video [9, 17, 25, 27, 38] cannot provide spatial localization. Furthermore, the latter schemes can only be used in an inefficient sliding window approach [38] for temporal localization.

Table 1. *Performance comparison on the UT-Interaction Dataset. The second and third columns report the recognition rate using the first half and the entire video, respectively. The annotation (Best) indicates the highest performance that the particular method can achieve using optimum parameters and visual vocabulary. (Avg.) indicates the average performance of the method using several different visual vocabularies.*

Method	Accuracy w. half videos	Accuracy w. full videos
Our Model	**73.3%**	**93.3%**
Ryoo [25] (Avg.)	61.8%	76.7%
Ryoo [25] (Best)	70%	85%
Cuboid+SVMs (Avg.) [9, 26]	25.3%	78%
Cuboid+SVMs (Best) [9, 26]	31.7%	85%
BP+SVM (Avg.) [25]	57.7%	75.9%
BP+SVM (Best) [25]	65%	83.3%
Yao *et al.* [37]	–	88%
Vahdat *et al.* [34]	–	93.3%
Zhang *et al.* [38]	–	**95%**
Kong *et al.* [14]	–	88.3%

Spatial Localization. As mentioned, our model is transparent and fully interpretable. By analyzing each keyframe's unary term and the pairwise terms, the most positively contributed poselets and pairs of poselets can be estimated. This analysis can lead to insightful information regarding the action: which keyframe is the most discriminative and which poselet or pair of poselets is the most distinctive. More specifically, the contribution of the presence ($\phi(x_{k_i})_j > 0$) or absence of the j-th poselet in the i-th keyframe can be easily identified as $w_{i_j}\phi(x_{k_i})_j$; a term of our scoring function (Eq. 1). Similarly, the contribution of the pairwise correlations between poselets of successive keyframes are computed. Fig. 4 shows visualization of this analysis for several test sequences. The bounding boxes of detected poselets that have most positive contributions to the scoring function of the action model are plotted. Further, the absent poselets that contribute positively via either the unary terms or the pairwise terms are annotated at the top right corner of each keyframe. Following this analysis, our model offers fine-level spatial localization of the action.

To evaluate the relevance of the selected detected poselets to the performed action, we define the localization score as $\frac{1}{|\mathbf{k}|}\sum_{i=1}^{|\mathbf{k}|}\frac{1}{P_i}\sum_{j=1}^{P_i}\frac{B_j \cap G_{k_i}}{B_j}$, where P_i is the total number of detected poselets with positive contributions in a keyframe, B_j is the corresponding bounding box, G_{k_i} is the ground truth bounding box enclosing the entire region of interest of the action. Considering each action model as a 1-vs-all classifier, we construct ROC curves based on the scores of each test sequence. A test sample is considered as positive prediction if its score is above a given threshold and its localization score above a threshold σ. The latter threshold defines the minimum average overlap of poselet's bounding boxes to consider them as part of the action. The average localization performance of all the action classes is shown in Fig. 5 (a). Our model achieves a 79% true positive rate at a 20% false positive rate with the threshold set to 0.5.



(a) (b)

Figure 3. (a) **Confusion** Matrix for UT-Interaction dataset. (b) Average classification accuracy for **different numbers** of keyframes.

6.2. Early Detection or Streaming

Our framework will not lose its discriminative power even if most of the frames of the test video are "dropped". The information that our model's scoring function extracts from a given image sequence is temporally sparse. Therefore, our framework is also suitable for human activity prediction in a streaming video scenario [25] without any specific modification. In order to investigate its quantitive performance on the latter task, we conducted experiments where the algorithm only had access to a fraction of the frames of the video. Similar to the experimental setting proposed by Ryoo [25], we tested our 4 keyframes action models using 10 different observation ratios. Fig. 5 (a) compares our model with existing methods. Our algorithm significantly outperforms the state-of-the-art method "Dynamic Cuboid BoW" [25] that was specifically developed for online activity prediction. Accessing only the first 70% of the test data frames, an average classification performance of 93.3% is acquired, which equals our performance observing the full videos (Sect. 6.1). Table 1 also lists the classification accuracy after having observed only the first half of the test video sequences.

6.3. Temporal Detection on Continuous Execution

We also examine the temporal localization ability of the proposed method. For this purpose, we focus our analysis on the unsegmented UT-Interaction sequences. The 10 videos have an average length of 3000 frames and contain executions of all 6 actions in random sequences while

Table 2. *Performance comparison of different components of our framework.*

Method	Our BoW+SVM	Volumetric Max-Pooled HoG+BoW Based Poselets+SVM	Unary Only HoG Based Poselets	Unary Only BoW Based Poselets	Unary Only HoG+BoW Based Poselets	Pairwise+Unary BoW Based Poselets	Pairwise+Unary HoG+BoW Based Poselets
Accuracy	73.3%	77.0%	82.0%	70.0%	90%	77.0%	**93.3**%

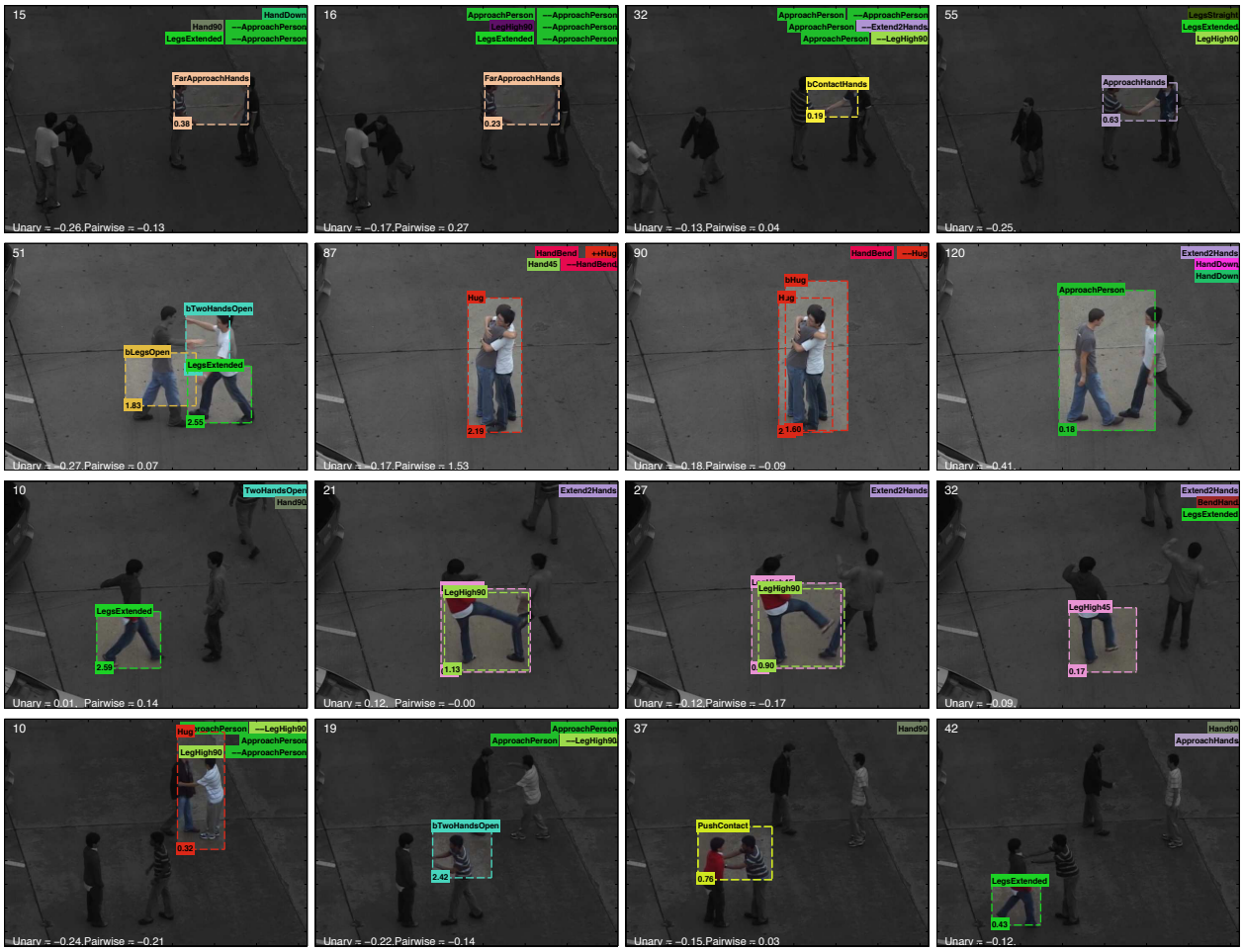
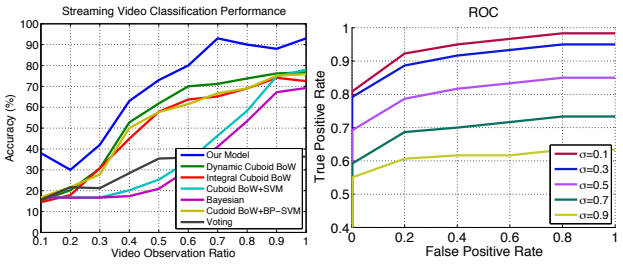


Figure 4. **Keyframe Interpretations.** Each row shows the 4 keyframes selected by our model from one of the test sequences; 4 out of a total 60 sequences illustrated. Detected or absent poselets and pairs of poselets contributing the most to our scoring function are automatically annotated. The bounding boxes of the detected poselets along with their name and score is plotted (the prefix 'b' points to BoW based detection). The ones absent and contributing positively are marked at the top left corner of each image. The pairs of poselet associated with our pairwise terms are also marked at the top right corner with their two corresponding strings. The first indicates the current keyframe poselet's name and the prefix ($++$ or $--$) of the second string indicates if this particular poselet is present or absent at the next keyframe. Based on those annotations an "action script" can be automatically created. For instance the action in the second row can be described: "*We first observe the poselets "LegExtended", "bLegsOpen" and "bTwoHandsOpen", followed by the poselet "Hug". A strong pairwise relationship appears due to absence of "HandBend" at the second keyframe and the presence of "Hug" at the third, etc.*"

also containing "background" activities. During our evaluation, an action class prediction is considered correct if it has a score above -1 and a percentage θ of the corresponding keyframes are contained inside the ground truth temporal segments. The average performance for all of the action classes is summarized in Table 3. We note that the 4 keyframes action models trained on the segmented videos (Sect. 6.1) were re-used in this experiment. During inference, the temporal distance between possible successive keyframes is constrained by the maximum temporal distance τ observed during the training phase (Sect. 4), *e.g.*, for the action "kicking" $\tau = 23$; for "hugging", $\tau = 55$. This constraint greatly improves the computation efficiency of our algorithm (linear with respect the number of frames) but also slightly boosts our detection performance. Overall our model attains a temporal detection accuracy of 80%, having predicted all the keyframes inside the correct temporal segment of the video (86.7% with a looser overlap threshold).



(a) (b)

Figure 5. (a) Classification accuracy with respect to the percentage of **streamed** frames observed. Our model (blue line) performs consistently above 80% after observing 60% of the video frames. The most important result, however, is that with less than half of the video observed, the classification performance is above 60%. (b) **Spatial Localization** of our model: ROC curves corresponding to 5 different localization score thresholds σ.

Table 3. *Detection on continuous execution videos.*

Temporal Overlap θ	0.25	0.5	0.75	1
Detection Accuracy	86.7%	86.7%	83.3%	80.0%

7. Discussion

We propose a new model for action recognition that combines a powerful mid-level representation, in the form of HoG and BoW poselets, with discriminative keyframe selection. The proposed approach has a number of important benefits, including the ability to spatially and temporally localize the action and deal with partial video observation (streaming). It also provides semantically interpretable output in the form of contextual temporal orderings of discriminant partial poses. In addition, the proposed model is easily extendable to incorporate more action parts, keyframes, or even generic objects for context.

References

[1] J. Assa, Y. Caspi, and D. Cohen-Or. Action synopsis: pose selection and illustration. *ACM Transactions on Graphics (TOG)*, 24, 2005. 1

[2] M. Blaschko and C. Lampert. Learning to localize objects with structured output regression. *ECCV*, 2008. 3, 4

[3] L. D. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE ICCV*, 2009. 1, 3, 4

[4] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *IEEE ICCV*, 2011. 2, 5

[5] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *IEEE Workshop on Models versus Exemplars in Computer Vision*, 2001. 2

[6] C. Chen and K. Grauman. Efficient activity detection with max-subgraph search. In *IEEE CVPR*, 2012. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005. 3

[8] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 2

[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005. 2, 6

[10] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE CVPR*, 2005. 2

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010. 5

[12] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *IEEE CVPR*, 2011. 2

[13] N. Ikizler and D. Forsyth. Searching video for complex activities with finite state models. *IEEE CVPR*, 2007. 2

[14] Y. Kong, Y. Jia, and Y. Fu. Learning human interaction by interactive phrases. *ECCV*, 2012. 3, 5, 6

[15] Z. Kourtzi and N. Kanwisher. Activation in human mt/mst by static images with implied motion. *J. of Cognitive Neuroscience*, 2000. 1

[16] C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE PAMI*, 31, 2009. 3

[17] I. Laptev. On space-time interest points. *IJCV*, 64, 2005. 2, 6

[18] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *NIPS*, 2010. 3

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE CVPR*, 2011. 3

[20] L. Liu, L. Shao, and P. Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 2012. 2

[21] S. Maji, L. D. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE CVPR*, 2011. 2, 3

[22] M. H. Nguyen and F. De la Torre. Max-margin early event detectors. In *IEEE CVPR*, 2012. 2

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

[24] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *IEEE CVPR*, 2012. 1, 2, 3

[25] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE ICCV*, 2011. 6

[26] M. Ryoo and J. Aggarwal. UT-Interaction dataset, ICPR contest on Semantic Description of human activities (SDHA), 2010. 5, 6

[27] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *IEEE CVPR*, 2012. 3, 5, 6

[28] S. Satkin and M. Hebert. Modeling the temporal extent of actions. *ECCV*, 2010. 2

[29] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? In *IEEE CVPR*, 2008. 2

[30] V. K. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *IEEE ICCV*, 2011. 1

[31] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *IEEE ICCV*, 2005. 2

[32] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *IEEE CVPR*, 2012. 1, 2

[33] C. Thurau and V. Hlavác. Pose primitive based human action recognition in videos or still images. In *IEEE CVPR*, 2008. 2

[34] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori. A discriminative key pose sequence model for recognizing human interactions. In *IEEE Int. Workshop on Visual Surveillance*, 2011. 1, 2, 3, 5, 6

[35] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *IEEE CVPR*, 2011. 1

[36] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *IEEE CVPR*, 2010. 2

[37] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *IEEE CVPR*, 2010. 6

[38] Y. Zhang, X. Liu, M. Chang, X. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *ECCV*, 2012. 5, 6

[39] Z. Zhao and A. M. Elgammal. Information theoretic key frame selection for action recognition. In *BMVC*, 2008. 2