Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification

Juan Carlos Niebles[1,2,3], Chih-Wei Chen[1], and Li Fei-Fei[1]

[1] Stanford University, Stanford CA 94305, USA
[2] Princeton University, Princeton NJ 08544, USA
[3] Universidad del Norte, Barranquilla, Colombia

Abstract. Much recent research in human activity recognition has focused on the problem of recognizing simple repetitive (walking, running, waving) and punctual actions (sitting up, opening a door, hugging). However, many interesting human activities are characterized by a complex temporal composition of simple actions. Automatic recognition of such complex actions can benefit from a good understanding of the temporal structures. We present in this paper a framework for modeling motion by exploiting the temporal structure of the human activities. In our framework, we represent activities as temporal compositions of motion segments. We train a discriminative model that encodes a temporal decomposition of video sequences, and appearance models for each motion segment. In recognition, a query video is matched to the model according to the learned appearances and motion segment decomposition. Classification is made based on the quality of matching between the motion segment classifiers and the temporal segments in the query sequence. To validate our approach, we introduce a new dataset of complex Olympic Sports activities. We show that our algorithm performs better than other state of the art methods.

Keywords: Activity recognition, discriminative classifiers.

1 Introduction

We argue that to understand motion, it is critical to incorporate temporal context information, particularly the temporal ordering of the movements. In this paper, we propose a simple discriminative framework for classifying human activities by aggregating information from motion segments that are considered both for their visual features as well as their temporal composition. An input video is automatically decomposed temporally into motion segments of variable lengths. The classifier selects a discriminative decomposition and combination of the segments for matching. Though simple in its form, we highlight a couple of advantages of our framework compared to the previous work.

First, depending on the time scale of the movement, actions have been traditionally grouped into: short but punctual actions (e.g. drink, hug), simple but periodic actions (e.g. walking, boxing), and more complex activities that are

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 392–405, 2010.

considered as a composition of shorter or simpler actions (e.g. a long jump, cooking). Very different algorithms have been proposed for these different types of motion, most of them take advantage of the special properties within its domain, hence perform rather poorly on other types. Our framework is a general one. No matter how simple or complex the motion is, our classifier relies on a temporal composition of various motion segments. Our basic philosophy is clear: temporal information helps action recognition at all time scales.

On the other hand, we note that some work has taken the approach of decomposing actions into "hidden states" that correspond to meaningful motion segments (i.e. HMM's, HCRF's, etc.). In contrast, we let the model automatically discover a robust combination of motion segments that improve the discriminability of the classifier. The result is a much simpler model that does not unnecessarily suffer from the difficult intermediate recognition step.

In order to test the efficacy of our method, we introduce a new dataset that focuses on complex motions in Olympic Sports, which can be difficult to discriminate without modeling the temporal structures. Our algorithm shows very promising results.

The rest of the paper is organized as follows. Section 1.1 overviews some of the related work. Section 2 describes a video representation that can be employed in conjunction with our model. Section 3 presents our model for capturing temporal structures in the data. We present experimental validation in Section 4 and conclude the paper in Section 5.

1.1 Related Work

A considerable amount of work has studied the recognition of human actions in video. Here we overview a few related work but refer the reader to [1,2] for a more complete survey.

A number of approaches have adopted the bag of spatio-temporal interest points [3] representation for human action recognition. This representation can be combined with either discriminative [4,5] classifiers, semi-latent topic models [6] or unsupervised generative [7,8] models. Such holistic representation of video sequences ignores temporal ordering and arrangement of features in the sequence.

Some researchers have studied the use of temporal structures for recognizing human activities. Methods based on dynamical Bayesian networks and Markov models have shown promise but either require manual design by experts [9] or detailed training data that can be expensive to collect [10]. Other work has aimed at constructing plausible temporal structures [11] in the actions of different agents but does not consider the temporal composition within the movements of a single subject, in part due to their holistic representation. On the other hand, discriminative models of temporal context have also being applied for classification of simple motions in rather simplified environments [12,13,14,15].

In addition to temporal structures, other contextual information can benefit activity recognition, such as background scene context [4] and object interactions [11,16]. Our paper focuses on incorporating temporal context, but does not exclude future work for combining more contextual information.

Our approach to capturing temporal structures is related to part-based models for object recognition. Both generative [17,18,19,20] and discriminative [21,22] models have shown promise in leveraging the spatial structures among parts for object recognition.

In this paper, we present a new representation for human activities in video. The key observation is that many activities can be described as a temporal composition of simple motion segments. At the global temporal level, we model the distinctive overall statistics of the activity. At shorter temporal ranges, we model the patterns in motion segments of shorter duration that are arranged temporally to compose the overall activity. Moreover, such temporal arrangement considered by our model is not rigid, instead it accounts for the uncertainty in the exact temporal location of each motion segment.

2 Video Representation

Our model of human actions can be applied over a variety of video descriptors. The key requirement is that a descriptor can be computed over multiple temporal scales, since our motion segment classifiers can operate on video segments of varying length. Frame-based representations and representations based on histograms are particular examples of descriptors that fit well to our framework. Here, we adopt a representation based on spatio-temporal interest points. Interest point based descriptors are attractive specially when tracking the subject performing the activity is difficult or not available. Several methods have been proposed for detecting spatio-temporal interest points in sequences [3,23,24]. In our approach, we use the 3-D Harris corner detector [3]. Each interest point is described by HoG (Histogram of Gradients) and HoF (Histogram of Flow) descriptors [5]. Furthermore, we vector quantize the descriptors by computing memberships with respect to a descriptor codebook, which is obtained by k-means clustering of the descriptors in the training set. During model learning and matching, we compute histograms of codebook memberships over particular temporal ranges of a given video, which are denoted by ψ_i in the following.

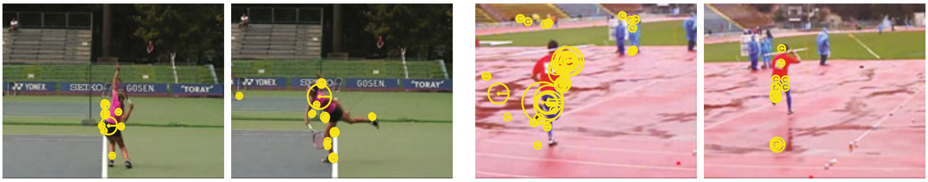

Fig. 1. Our framework can be applied over a variety of video data representations. Here we adopt a representation based on spatio-temporal interest points. This figure shows example spatio-temporal interest points detected with the 3D Harris corner method from [3]. Video patches are extracted around each point, and described by their local shape and motion patterns.

3 Modeling Temporal Structures

In this section we present our framework for recognizing complex human activities in video. We propose a temporal model for recognizing human actions that incorporates simple motion segment classifiers of multiple temporal scales. Fig. 2 shows a schematic illustration of our human action model. The basic philosophy is very simple: a video sequence is first decomposed into many temporal segments of variable length (including the degenerate case of the full sequence itself). Each video segment is matched against one of the motion segment classifiers by measuring image-based similarities as well as the temporal location of the segment with respect to the full sequence. The best matching scores from each motion segment classifier are accumulated to obtain a measure of the matching quality between the full action model and the query video. As Fig. 2 illustrates, an action model encodes motion information at multiple temporal scales. It also encodes the ordering in which the motion segments tend to appear in the sequence. In the following, we discuss the details of the model, the recognition process and learning algorithm.

3.1 Model Description

Here we introduce the model of human actions, which is illustrated in Fig. 2. Our full action model is composed by a set of K motion segment classifiers $A_1, ..., A_K$, each of them operating at a particular temporal scale. Each motion segment classifier A_i operates over a histogram of quantized interest points extracted from a temporal segment whose length is defined by the classifier's temporal scale s_i. In addition to the temporal scale, each motion segment classifier also specifies a temporal location centered at its preferred anchor point t_i. Lastly, the motion segment classifier is enriched with a flexible displacement model τ_i that captures the variability in the exact placement of the motion segment A_i within the sequence.

We summarize the parameters of our model with the parameter vector $\mathbf{w}$ as the concatenation of the motion segment classifiers and the temporal displacement parameters,

$$\mathbf{w} = (A_1, ..., A_K, \tau_1, ..., \tau_K). \tag{1}$$

3.2 Model Properties

Our model addresses the need to consider temporal structure in the task of human activity classification. In the following, we discuss some important properties of our framework.

Coarse-to-fine motion segment classifiers. Our model contains multiple classifiers at different time scales, enabling it to capture characteristic motions of various temporal granularity. On one end, holistic bag-of-features operate at the coarsest scale, while frame-based methods operate at the finest scale. Our

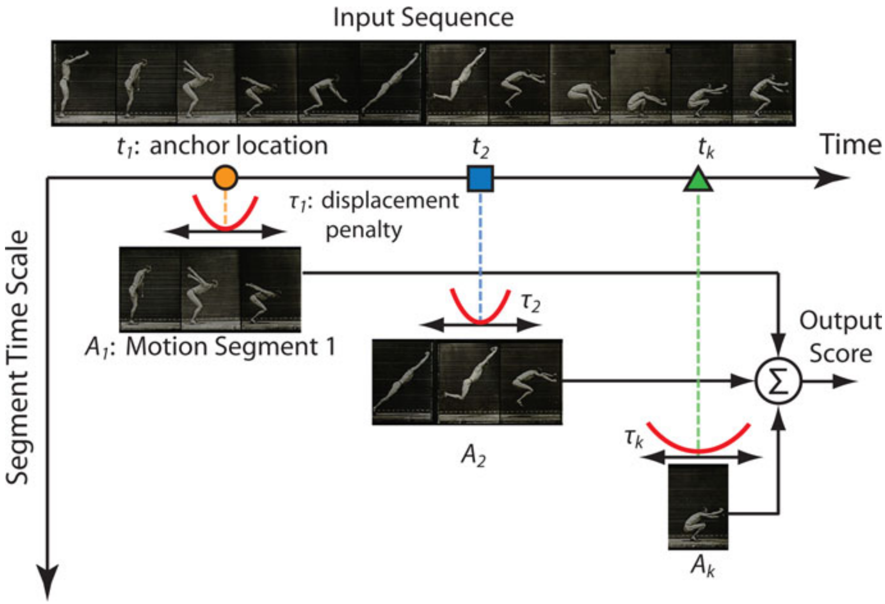

Fig. 2. Model Architecture. Here we show the structure of our model for activity recognition. The input video V is described by histograms of vector quantized interest points, which are computed over multiple temporal ranges. Each motion segment classifier A_i has a particular temporal scale, and it is matched to the features $\psi_i(V, h_i)$ from temporal segments of the input sequence of that temporal extent. The optimal location of each motion segment classifier is determined by the appearance similarity $(A_i \cdot \psi_i(V, h_i))$ and penalty of temporal displacement from the anchor point t_i $(\tau_i \cdot \psi_(h_i - t_i))$. The overall matching score combines scores of individual components. A classification decision is made by thresholding the resulting matching score. See Sec. 3 for more details.

framework has the flexibility to operate between these two ends of the temporal spectrum, and it closes the gap by allowing multiple classifiers to reside in a continuum of temporal scales.

Temporal Context. While discriminative appearance is captured by our multiple classifiers at different time scales, the location and order in which the motion segments occur in the overall activity also offer rich information about the activity itself. Our framework is able to capture such temporal context: the anchor points of the motion segment classifiers encode the temporal structure of the activity. In particular, these canonical positions prohibit the classifiers from matching time segments that are distant from them. This implicitly carries ordering constraints that are useful for discriminating human activities.

Flexible Model. Equipped with classifiers of multiple time scales and the temporal structure embedded in their anchor points, our model is capable of searching for a best match in a query sequence and score it accordingly. However, the temporal structure in videos of the same class might not be perfectly aligned. To handle intra-class variance, our model incorporates a temporal displacement penalty that allows the optimal placement of the each motion segment to deviate from its anchor point.

3.3 Recognition

Given a trained model, the task in recognition is to find the best matching of the model to an input sequence. This requires finding the best scoring placement for each of the K motion segment classifiers. We denote a particular placement of the motion segment classifiers within a sequence V by a hypothesis $H = (h_1, ..., h_k)$. Each h_i defines the temporal position for the i-th motion segment classifier. We measure the matching quality of motion segment classifier A_i at location h_i by favoring good appearance similarity between the motion segment classifier and the video features, and penalizing for the temporal misplacement of the motion segment classifier when h_i is far from the anchor point t_i. That is, the matching score for the i-th motion segment classifier is

$$A_i \cdot \psi_i(V, h_i) - \tau_i \cdot \psi_{di}(h_i - t_i). \tag{2}$$

In the first term of Eq. 2, which captures the appearance similarity, $\psi_i(V, h_i)$ is the appearance feature vector (i.e. histogram of quantized interest points) extracted at location h_i with scale s_i. In our experiments, we implement the classifier A_i with a χ^2 support vector machine. The kernel function for A_i is given by

$$K(x_k, x_j) = \exp\left(-\frac{1}{2S}\sum_{r=1}^{D}\frac{(x_{kr} - x_{jr})^2}{x_{kr} + x_{jr}}\right), \tag{3}$$

where S denotes the mean distance among training examples, $\{x_{ki}\}_{i=1...D}$ are the elements of the histogram x_k and D is the histogram dimensionality. In practice, D is equal to the size of the codebook. In the second term of Eq. 2, which captures the temporal misplacement penalty, $\psi_{di}(h_i - t_i)$ denotes the displacement feature. The penalty, parametrized by $\tau_i = \{\alpha_i, \beta_i\}$, is a quadratic function of the motion segment displacement and given by

$$\tau_i \cdot \psi_{di}(h_i - t_i) = \alpha_i \cdot (h_i - t_i)^2 + \beta_i \cdot (h_i - t_i). \tag{4}$$

We obtain an overall matching score for hypothesis H by accumulating the scores from all motion segment classifiers in the model:

$$\sum_{i=1}^{K} A_i \cdot \psi_i(V, h_i) - \tau_i \cdot \psi_{di}(h_i - t_i). \tag{5}$$

Let $f_{\mathbf{w}}(V)$ be a scoring function that evaluates sequence V. In recognition, we consider all possible hypotheses and choose the one with the best matching score:

$$f_{\mathbf{w}}(V) = \max_{H} \sum_{i=1}^{K} A_i \cdot \psi_i(V, h_i) - \tau_i \cdot \psi_{di}(h_i - t_i). \tag{6}$$

A binary classification decision for input video V is done by thresholding the matching score $f_{\mathbf{w}}(V)$.

There is a large number of hypotheses for a given input video sequence. However, note that once the appearance similarities between the video sequence and each motion segment classifier are computed, selecting the hypothesis with the best matching score can be done efficiently using dynamic programming and distance transform techniques [18] in a similar fashion to [21,25].

3.4 Learning

Suppose we are given a set of example sequences $\{V^1, \ldots, V^N\}$ and their corresponding class labels $y_{1:N}$, with $y_i \in \{1, -1\}$. Our goal is to use the training examples to learn the model parameters $\mathbf{w}$. This can be formulated as the minimization of a discriminative cost function. In particular, we consider the following minimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i f_{\mathbf{w}}(V^i)), \tag{7}$$

where C controls the relative weight of the hinge loss term. This is the formulation of a Latent Support Vector Machine (LSVM) [21]. In the LSVM framework, the scoring function maximizes over the hidden variables. In our method, the hidden variables correspond to the best locations of the motion segment classifiers on each training video. Note that it is not necessary to supervise the locations of the motion segment classifiers during training, instead this is a weakly supervised setting, where only a class label is provided for each example.

The optimization problem described above is, in general, non-convex. However, it has been shown in [21] that the objective function is convex for the negative examples, and also convex for the positive examples when the hidden variables are fixed.

This leads to an iterative learning algorithm that alternates between estimating model parameters and estimating the hidden variables for the positive training examples. In summary the procedure is as follows. In the first step, the model parameters $\mathbf{w}$ are fixed. The best scoring locations $H_p^\star$ of the motion segment classifiers are selected for each positive example p. This is achieved by running the matching process described in Section 3.3 on the positive videos. In the second step, by fixing the hidden variables of the positive examples to the locations given by $H_p^\star$, the optimization problem in Eq. 7 becomes convex. We select negative examples by running the matching process in all negative training videos and retrieving all hypotheses with large matching score. We train the parameters $\mathbf{w}$ using LIBSVM [26] on the resulting positive and negative examples. This process is repeated for a fixed small number of iterations.

In most cases, the iterative algorithm described above requires careful initialization. We choose a simple initialization heuristic. First, we train a classifier

Table 1. Left: Accuracy for action classification in the KTH dataset. Right: Comparison of our model to current state of the art methods.

Action Class	Our Model
walking	94.4%
running	79.5%
jogging	78.2%
hand-waving	99.9%
hand-clapping	96.5%
boxing	99.2%

Algorithm	Perf.
Ours	91.3%
Wang et al. [28]	92.1%
Laptev et al. [5]	91.8%
Wong et al. [8]	86.7%
Schuldt et al. [27]	71.5%
Kim et al. [29]	95%

with a single motion segment classifier that covers the entire sequence. This is equivalent to training a χ^2-SVM on a holistic bag of features representation. We then augment the model with the remaining $K-1$ motion segment classifiers. The location and scale of each additional motion segment classifier is selected so that it covers a temporal range that correlates well with the global motion segment classifier. This favors temporal segments that exhibit features important for overall discrimination.

4 Experimental Results

In order to test our framework, we consider three experimental scenarios. First, we test the ability of our approach to discriminate simple actions on a benchmark dataset. Second, we test the effectiveness of our model at leveraging the temporal structure in human actions on a set of synthesized complex actions. Last, we present a new challenging Olympic Sports Dataset and show promising classification results with our method.

4.1 Simple Actions

We use the KTH Human actions dataset [27] to test the ability of our method to classify simple motions. The dataset contains 6 actions performed by 25 actors, for a total of 2396 sequences. We follow the experimental settings described in [27]. In all experiments, we adopt a representation based on spatio-temporal interest points described by concatenated HoG/HoF descriptors. We construct a codebook of local spatio-temporal patches from feature descriptors in the training set. We set the number of codewords to be 1000. Experimental results are shown in Table 1. A direct comparison is possible to the methods that follow the same experimental setup [5,8,27,28]. We note that our method shows competitive results, but its classification accuracy is slightly lower than the best result reported in [28].

4.2 Synthesized Complex Actions

In this experiment, we aim to test the ability of our model to leverage the temporal structure of human actions. In order to test this property in a controlled

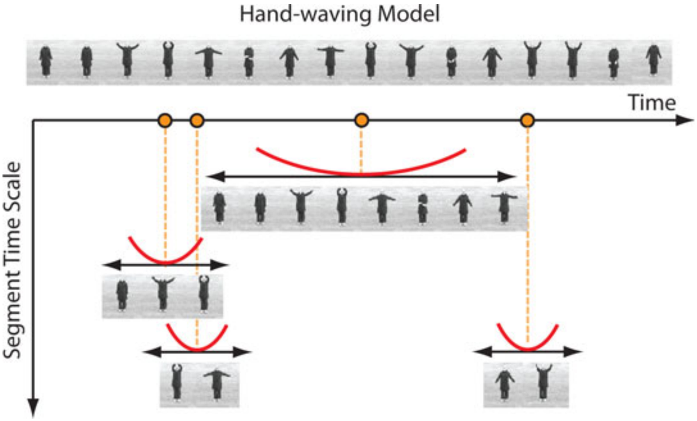

Fig. 3. An example of our learned model. In this illustration, the horizontal axis represents time. Each row corresponds to a motion segment classifier learned by our model whose temporal extent is indicated by its vertical location. The appearance of the motion segment is illustrated by a few example frames. The associated dot indicates the anchor position t_i of the motion segment relative to the full sequence. The parameters of the temporal misplacement penalty τ_i are represented by the parabola centered at the anchor point. Notice that the vertical arrangement of the motion segments shows the distinct temporal scales at which each classifier operates.

setting, we construct a synthesized set of complex actions by concatenating 3 simple motions from the Weizmann action database: 'jump', 'wave' and 'jack'. In total, we synthesize 6 complex actions classes by concatenating one video of each simple motion into a long sequence.

In this test, a baseline model that uses a single motion segment classifier covering the entire video sequence performs at random chance or ≈ 17%. The simple holistic bag-of-features has trouble differentiating actions in this set since the overall statistics are nearly identical. On the other hand, our model which takes advantage of temporal structure and orderings, can easily discriminate the 6 classes and achieve perfect classification performance at 100%. In Fig. 4 we show a learned model for the complex action composed by 'wave'-'jump'-'jack'. Notice that our model nicely captures discriminative motion segments such as the transitions between 'jump' and 'jack'.

4.3 Complex Activities: Olympic Sports Dataset

We have collected a dataset of Olympic Sports activities from YouTube sequences. Our dataset contains 16 sport classes, with 50 sequences per class. See Fig. 5 for example frames from the dataset. The sport activities depicted in the dataset contain complex motions that go beyond simple punctual or repetitive actions[1]. For instance, sequences from the long-jump action class, show an

[1] In contrast to other sport datasets such as [15], which contains periodic or simple actions such as walking, running, golf-swing, ball-kicking.

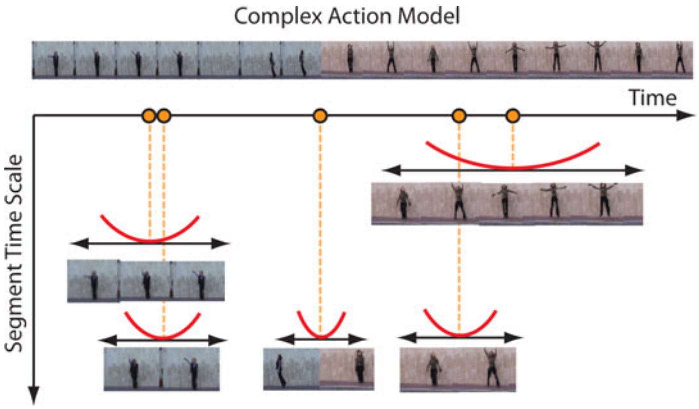

Fig. 4. A learned model for the synthesized complex action ‘wave’-‘jump’-‘jack’. See Fig. 3 for a description of the illustration.

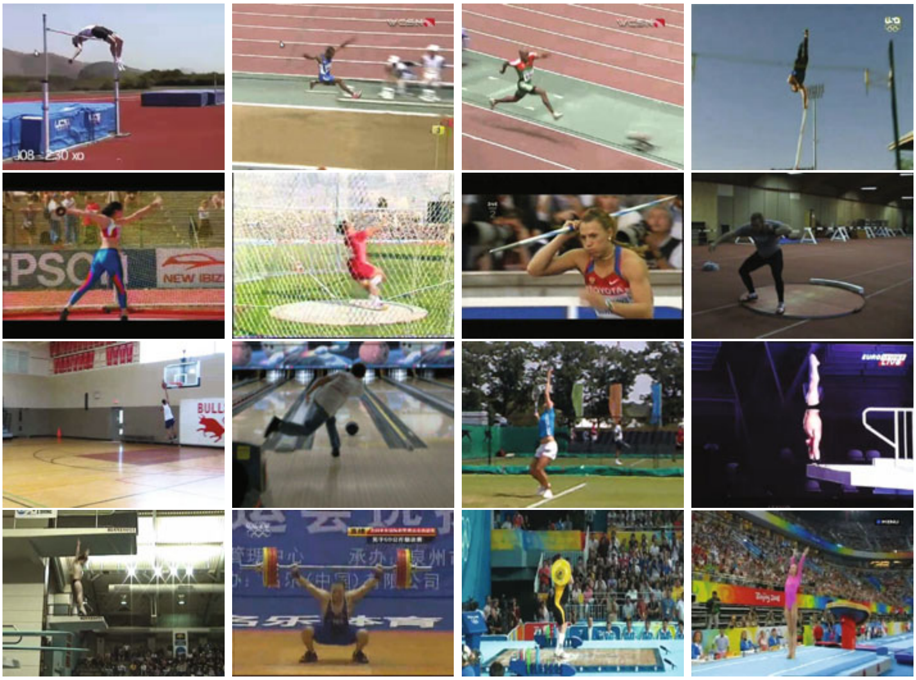

Fig. 5. Olympic Sports Dataset. Our dataset contains 50 videos from each of 16 classes: high jump, long jump, triple jump, pole vault, discus throw, hammer throw, javelin throw, shot put, basketball layup, bowling, tennis serve, platform (diving), springboard (diving), snatch (weightlifting), clean and jerk (weightlifting) and vault (gymnastics). The sequences, obtained from YouTube, contain severe occlusions, camera movements, compression artifacts, etc. The dataset is available at `http://vision.stanford.edu`.

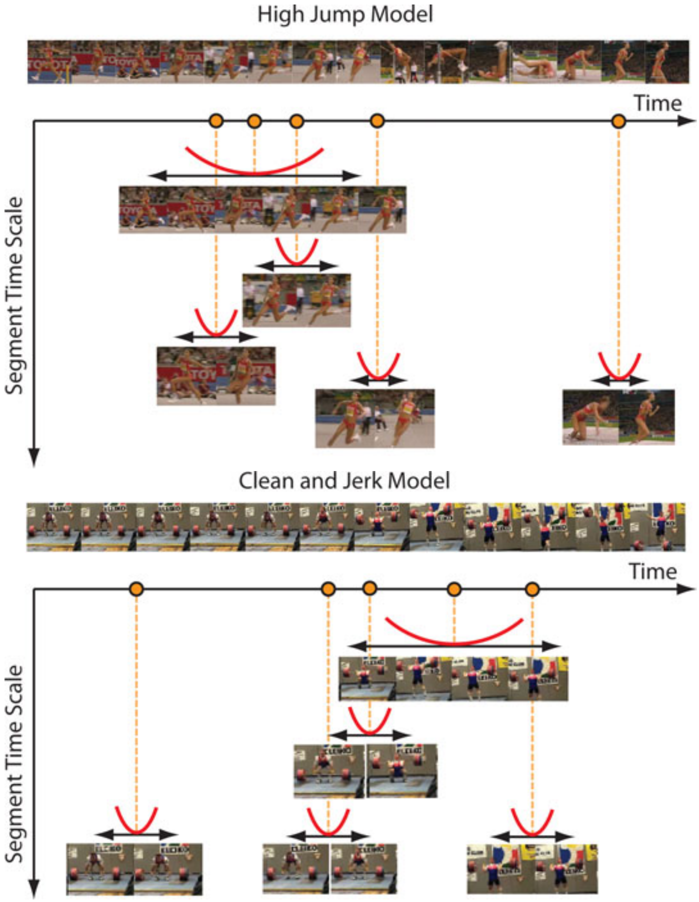

Fig. 6. Learned model for the complex actions in the Olympic Sports Dataset: high-jump and clean-and-jerk. See Fig. 3 for a description of the illustration.

athlete first standing still, in preparation for his/her jump, followed by running, jumping, landing and finally standing up. The dataset is available for download at our website `http://vision.stanford.edu`.

We split the videos from each class in the dataset into 40 sequences for training and 10 for testing. We illustrate two of the learned models in Fig. 6. Table 2 shows the classification results of our algorithm. We compare the performance of our model to the multi-channel method of [5], which incorporates rigid spatio-temporal binnings and captures a rough temporal ordering of features.

Finally, Fig. 7 shows three learned models of actions in the Olympic Sports dataset, along with matchings to some testing sequences. In the long jump example, the first motion segment classifier covers the running motion at the beginning of the sequence. This motion segment has a low displacement penalty over a large temporal range as indicated by its wide parabola. It suggests that the model has

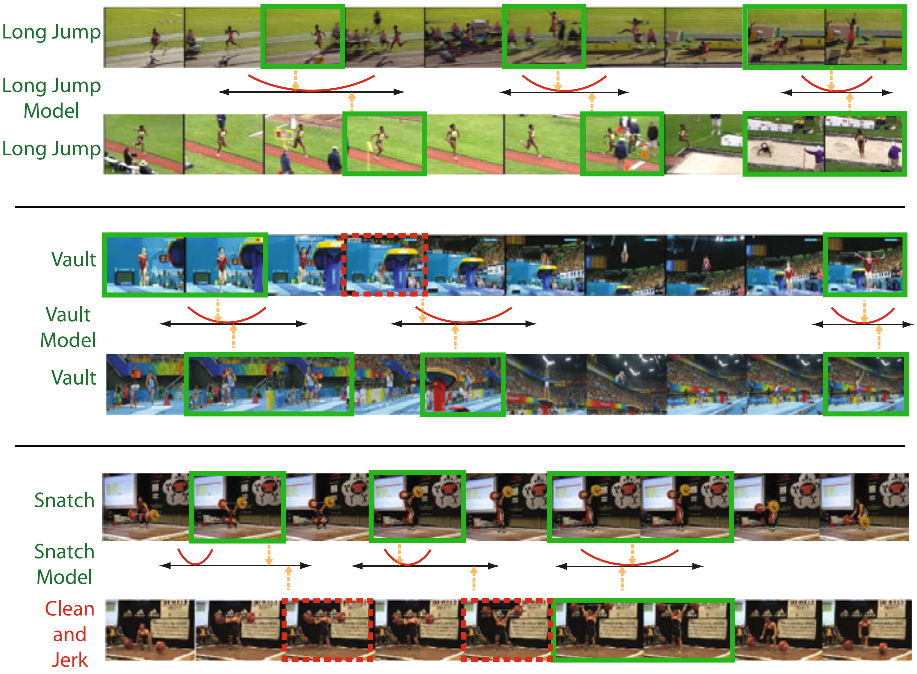

Fig. 7. We illustrate learned action models for long jump, vault and snatch. Each group depicts two testing sequences (top and bottom), as well as an illustration of the temporal displacement penalty parameters (middle). Green boxes surround matched temporal segments that are most compatible with the corresponding motion segment classifiers. Red boxes indicate temporal segments that are matched to the motion segment model with a low matching score. The arrows indicate the automatically selected best placement for each motion segment.

Table 2. Average Precision (AP) values for the classification task in our Olympic Sports Dataset

Sport class	Our Method	Laptev et al. [5]	Sport class	Our Method	Laptev et al. [5]
high-jump	**68.9%**	52.4%	javelin-throw	**74.6%**	61.1%
long-jump	**74.8%**	66.8%	hammer-throw	**77.5%**	65.1%
triple-jump	**52.3%**	36.1%	discus-throw	**58.5%**	37.4%
pole-vault	**82.0%**	47.8%	diving-platform	87.2%	**91.5%**
gymnastics-vault	86.1%	**88.6%**	diving-springboard	77.2%	**80.7%**
shot-put	**62.1%**	56.2%	basketball-layup	**77.9%**	75.8%
snatch	**69.2%**	41.8%	bowling	**72.7%**	66.7%
clean-jerk	**84.1%**	83.2%	tennis-serve	**49.1%**	39.6%
			Average (AAP)	**72.1%**	62.0%

learned to tolerate large displacements in the running stage of this activity. On the other hand, in the vault example, the middle motion segment classifier has a low matching score to the top testing sequence. However, the matching scores in other temporal segments are high, which provides enough evidence to the full action model for classifying this sequence correctly. Similarly, the bottom clean and jerk sequence in the snatch model obtains a high matching score for the last motion segment, but the evidence from the motion segments is rather low. We also observe that our learned motion segment classifiers display a wide range of temporal scales, indicating that our model is able to capture characteristic motion patterns at multiple scales. For example, the longer segments that contain the athlete holding the weights in the snatch model, and the shorter segments that enclose a jumping person in the long jump model.

5 Conclusion and Future Work

In this paper we have empirically shown that incorporating temporal structures is beneficial for recognizing both complex human activities as well as simple actions. Future directions include incorporating other types of contextual information and richer video representations.

Acknowledgments. This project is partially supported by an NSF CAREER grant and a Kodak Award to L.F-F. We would like to thank Barry Chai, Olga Russakovsky, Hao Su, Bangpeng Yao and the anonymous reviewers for their useful comments.

References

1. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine Recognition of Human Activities: A Survey. IEEE Transactions on Circuits and Systems for Video Technology 18, 1473–1488 (2008)
2. Forsyth, D.A., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D.: Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. Foundations and Trends in Computer Graphics and Vision 1, 77–254 (2005)
3. Laptev, I.: On Space-Time Interest Points. IJCV 64, 107–123 (2005)
4. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR, pp. 2929–2936. IEEE, Los Alamitos (2009)
5. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, p. 18. IEEE, Los Alamitos (2008)
6. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE TPAMI 31, 1762–1774 (2009)
7. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. IJCV 79, 299–318 (2008)
8. Wong, S.F., Kim, T.K., Cipolla, R.: Learning Motion Categories using both Semantic and Structural Information. In: CVPR, pp. 1–6. IEEE, Los Alamitos (2007)
9. Laxton, B., Lim, J., Kriegman, D.: Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In: CVPR. IEEE, Los Alamitos (2007)

10. Ikizler, N., Forsyth, D.A.: Searching for Complex Human Activities with No Visual Examples. IJCV 80, 337–357 (2008)
11. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: CVPR, pp. 2012–2019. IEEE, Los Alamitos (2009)
12. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. CVIU 104, 210–220 (2006)
13. Wang, S.B., Quattoni, A., Morency, L.P., Demirdjian, D., Darrell, T.: Hidden Conditional Random Fields for Gesture Recognition. In: CVPR, vol. 2, pp. 1521–1527. IEEE, Los Alamitos (2006)
14. Quattoni, A., Wang, S.B., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE TPAMI 29, 1848–1853 (2007)
15. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In: CVPR. IEEE, Los Alamitos (2008)
16. Yao, B., Fei-Fei, L.: Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In: CVPR. IEEE, Los Alamitos (2010)
17. Bouchard, G., Triggs, B.: Hierarchical Part-Based Visual Object Categorization. In: CVPR, pp. 710–715. IEEE, Los Alamitos (2005)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. IJCV 61, 55–79 (2005)
19. Fergus, R., Perona, P., Zisserman, A.: Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition. IJCV 71, 273–303 (2007)
20. Niebles, J.C., Fei-Fei, L.: A Hierarchical Model of Shape and Appearance for Human Action Classification. In: CVPR, pp. 1–8. IEEE, Los Alamitos (2007)
21. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. IEEE TPAMI, 1–20 (2009)
22. Ke, Y., Sukthankar, R., Hebert, M.: Event Detection in Crowded Videos. In: ICCV, pp. 1–8. IEEE, Los Alamitos (2007)
23. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: VSPETS, pp. 65–72. IEEE, Los Alamitos (2005)
24. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. In: ICCV, vol. 2, pp. 1395–1402. IEEE, Los Alamitos (2005)
25. Felzenszwalb, P.F., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR, pp. 1–8. IEEE, Los Alamitos (2008)
26. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
27. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR, pp. 32–36. IEEE, Los Alamitos (2004)
28. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)
29. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor Canonical Correlation Analysis for Action Classification. In: CVPR, pp. 1–8. IEEE, Los Alamitos (2007)