A Hierarchical Representation for Future Action Prediction

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese

Stanford University, USA

Abstract. We consider inferring the future actions of people from a still image or a short video clip. Predicting future actions before they are actually executed is a critical ingredient for enabling us to effectively interact with other humans on a daily basis. However, challenges are two fold: First, we need to capture the subtle details inherent in human movements that may imply a future action; second, predictions usually should be carried out as quickly as possible in the social world, when limited prior observations are available.

In this paper, we propose *hierarchical movemes* - a new representation to describe human movements at multiple levels of granularities, ranging from atomic movements (e.g. an open arm) to coarser movements that cover a larger temporal extent. We develop a max-margin learning framework for future action prediction, integrating a collection of moveme detectors in a hierarchical way. We validate our method on two publicly available datasets and show that it achieves very promising performance.

1 Introduction

Every day, humans are faced with numerous situations in which they must predict what actions other people are about to do in the near future. These predictions are a critical ingredient for enabling us to effectively interact with other humans on a daily basis. Consider the example shown in Fig. 1. When presented with a short video clip or even a static image, we can easily predict what is going to happen in the near future (e.g. the man and the woman are about to hug). The ability of the human visual system to predict future actions is possibly thanks to years of previous observations of interactions among humans.

Predicting the action of a person before it is actually executed has a wide range of applications in autonomous robots, surveillance and health care. For autonomous navigation, in order for an agent to safely and effectively operate alongside humans, it is important for it to predict what people are about to do next. This ability can enable the robot to plan ahead for reactive responses or to avoid potential accidents. For example, if an autonomous agent observes a person that is losing balance, then it is highly probable that s/he would fall. If the vehicle can predict it, then it would stop and thus avoid an accident.

In this paper, we consider the problem of future action prediction in natural (non-staged) scenarios. Given a large collection of training videos containing

D. Fleet et al. (Eds.): ECCV 2014, Part III, LNCS 8691, pp. 689–704, 2014.

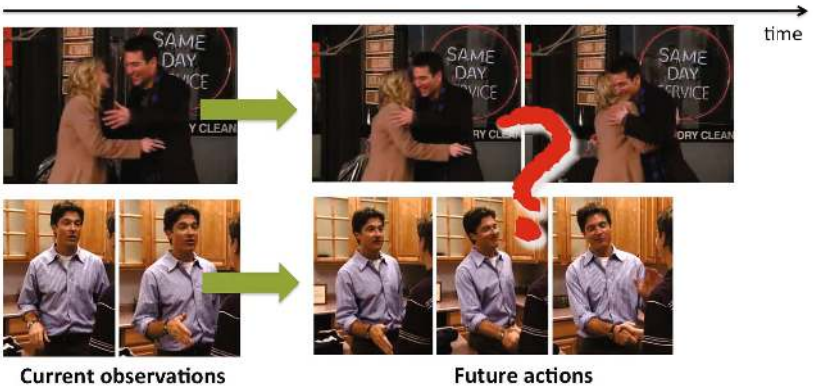


Fig. 1. Future action prediction. Given a static image or a short video clip (left), our goal is to infer the actions (as well as a sequence of movements) that are going to happen in the near future. The key contribution of this paper is to unveil the subtle details behind these movements and make correct action predictions.

human actions in the real world (e.g. TV series), we learn how human behaviors tend to evolve dynamically in a short period of time. Our goal is to infer the action that a person is going to perform next, from the observation of a short video clip or even a single frame.

Compared to the well-studied human action recognition, there are two characteristics of future action prediction: First, predicting future actions requires identifying the fine-grained details inherent to the current observations that would lead to a future action. For example, seeing a person with open arms indicates that s/he is probably going to hug. Second, it is often the case that future action prediction must be carried out with only the short-term observations of people in a short video clip or even a static image. It is important for an autonomous robot to react to the environments (e.g. a person appearing unexpectedly) as quickly as possible.

This paper introduces a new representation called *hierarchical movemes*, which is able to capture the typical structure of human movements before an action is executed. The term *"moveme"* was first introduced in the early work of Bregler [1], which is used to represent the atomic component of human movements, such as reaching and grabbing [1,5]. We generalize the notion of "movemes" to capturing human movements at multiple levels of semantic and temporal granularity, ranging from an atomic motion with consistent viewpoints lasting a few frames, to larger motion segments covering more than one atomic motion. In the extreme case, we have "movemes" depicting all possible movements prior to an action. An example of hierarchical movemes representation is shown in Fig. 2. Given a new image or a short video clip, we infer the action that is going to happen using this hierarchical representation. In this paper, we focus on modeling human movements before the action is actually executed.

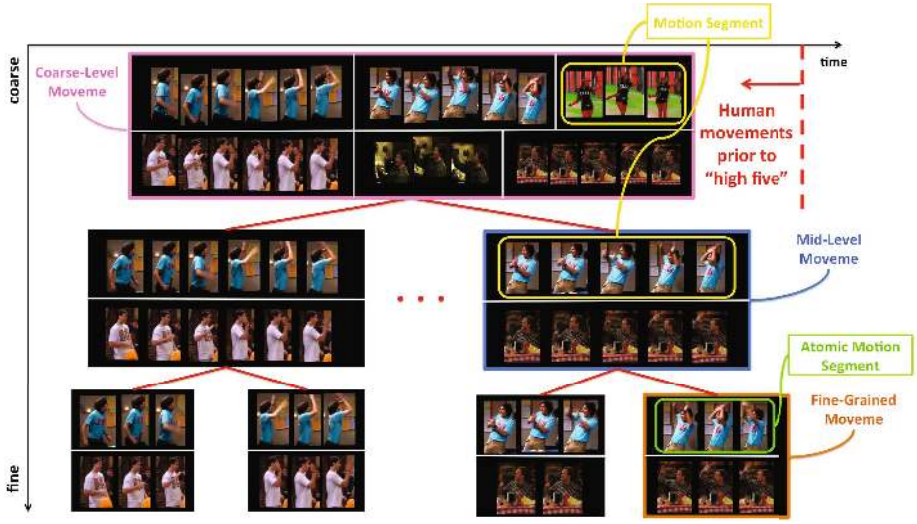


Fig. 2. An illustration of the *hierarchical movemes* representation. In this example, the structure of human movements prior to "high five" is represented, from coarse to fine. At the top level, movemes (*coarse-level movemes*) capture generic viewpoint and pose characteristics of the future action we wish to predict (i.e. that take place before the action we want to predict). At the second level, movemes (*mid-level movemes*) capture viewpoint-specific but pose-generic characteristics of the future action. At the lower level, movemes (*fine-grained movemes*) capture viewpoint-specific AND pose-specific characteristics of the future action.

However, the representation is general and can be applied to recognizing observed actions with complex structures.

2 Previous Work

Human action recognition is an extremely important research area in computer vision, and has grown dramatically in the last decade. Recent research has stepped past recognizing simple human actions, such as walking and standing in constrained settings [19], and gradually moved towards understanding complex actions in realistic video and still images collected from movies [11], TV shows [14], sport games [10], internet [28], etc. These scenarios typically include background clutter, occlusions, viewpoint changes, etc and have imposed significant challenges on action recognition. In the video domain, bag-of-features representations of local space-time features [22] have achieved impressive results. In the image domain, the contextual information such as attributes [28], objects [27] and poses [25,23] are jointly modeled with actions.

Recent research in early event detection has attempted to expand the spectrum of human action recognition to actions in the future. Ryoo [18] addresses the problem of early recognition of unfinished activities. Two variants of the

bag-of-words representations are introduced to handle the computational issues of modeling how feature distributions change over time. Hoai and Torre [6] introduces a structural SVM framework for early event detection. A slack-rescaling approach is proposed to constrain the monotonicity among past, partial, complete and future event. Our work differs from previous literatures on early event detection in three aspects: 1) Our method is able to predict future actions from any timestamp in a video. This is in sharp contrast to the early event detection approaches that constrain the input to the "early stage of an action". 2) Previous works typically require relatively long prior observations of actions, our method can predict from a short video clip or even a static image. 3) We expand the scope of action prediction from controlled lab settings (as in [18] and [6]) to unconstrained "in-the-wild" footage.

The importance of future action prediction has been demonstrated recently in robotic applications [24,9]. For example, Koppula and Saxena [9] address the problem of anticipating future activities from RGB-D data by considering human-object interactions. The method has been implemented into a real robotic system to assist humans in daily tasks such as opening the fridge door and refilling water glasses.

Predicting the future exists in other domains of computer vision. Most of the works are focused on predicting (or forecasting) the future trajectories of pedestrians [15,7]. There are also literatures on predicting motion from still images [29]. Our work is philosophically similar to these, but we focus on predicting motion patterns associated to semantically meaningful actions..

We highlight the main contributions of our paper. 1) We consider predicting future actions from still images or short video clips in unconstrained data. There is a body of work [18,6] that considers early action prediction from stream videos in constrained settings. This paper is the first that attempts to predict future actions from a single frame in the challenging real-world scenarios. 2) We introduce a novel representation called *hierarchical movemes* to capture multiple levels of granularities in human movements. 3) We develop a max-margin learning framework that jointly learns the appearance models of different movemes, as well as their relations. We demonstrate experimentally that this framework is effective in future action prediction.

3 Hierarchical Movemes - A New Representation for Actions in the Future

Modeling human actions is a very challenging problem in that: 1) Humans are highly articulated objects; 2) Actions can be described at different levels of semantic granularities, ranging from higher level actions, such as handshaking and talking, to finer grained motions, such as reaching and grabbing. Traditional action recognition methods usually focus on recognizing the higher level action classes. In action prediction, however, critical clues are usually hidden in finer grained motions. For example, an open arm usually implies hugging, but "open arm" is not necessarily an important class for action recognition.

In this paper, we propose a new representation called *hierarchical movemes* for future action prediction. The hierarchy depicts human movements at multiple levels of granularities from coarse to fine. An example of hierarchical moveme representation is shown in Fig. 2. We start by describing the procedure of constructing the hierarchy.

3.1 Hierarchy Construction

During training, we assume that we are given a collection of videos annotated with bounding boxes around the true locations of the people in each frame, tracks associated with each person across frames, action and viewpoint labels for each frame. We use the tracks associated with each person in the training videos to construct the hierarchy. We truncate the tracks that contain an action of interest (e.g. handshake, hug, kiss), such that the last frame of each track is right before the starting point of the action we want to predict. This allows our learning algorithm to only focus on modeling people's movements before actions are executed. See Fig. 3 for an example.

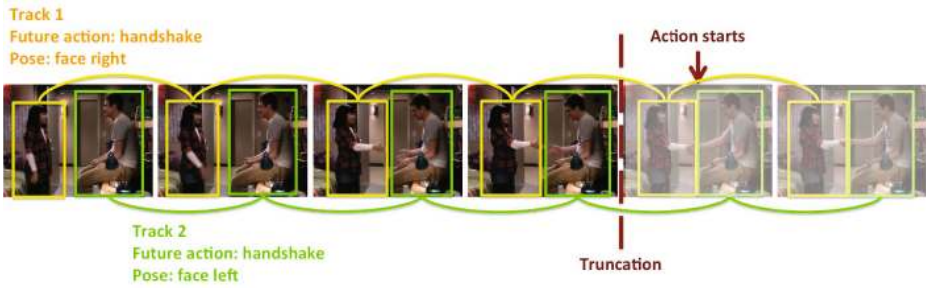


Fig. 3. Example of annotations for training [14]. Annotations include bounding boxes around the true locations of the people in each frame, tracks associated with each person across frames, action and viewpoint labels for each frame. We truncate the tracks associated with each person, such that the last frame of each track is right before the starting point of the action we want to predict. We define the starting point of an action according to the annotation of the dataset [14]. For example, persons are labeled as "handshake" when their hands touch each other.

We construct a 3-layer "moveme" hierarchy to capture human movements at different levels of semantic and temporal granularity. An example hierarchy is shown in Fig. 2. At the top level, movemes (which we call *coarse-level movemes*) capture generic viewpoint and pose characteristics of the future action we wish to predict (using frames that take place before the action we want to predict). For example, in Fig. 2, this layer captures a collection of generic human movements that lead to the action "high five" in the near future. At the second level, movemes (which we call *mid-level movemes*) capture viewpoint-specific but pose-generic characteristics of the future action. For example, Fig. 2 shows

two movemes in the second layer that are associated to movements observed from the "right" viewpoint, the other from the "left" respectively. At the lower level, movemes (which we call *fine-grained movemes*) capture viewpoint-specific and pose-specific characteristics of the future action. For example, in Fig. 2, the second fine-grained moveme in the third layer represents movements observed from the right and correspond to a pose configuration where arms are raised.

In training, the labels (including actions and viewpoints) for the coarse-level and mid-level movemes are given, while the fine-grained movemes are automatically discovered from the training data. In the following, we will introduce how to discover the fine-grained movemes via discriminative temporal clustering. An overview of the clustering process is shown in Fig. 4.

Fine-Grained Moveme Discovery. Given "mid-level movemes" that correspond to movements of people with consistent viewpoints, our goal is to partition the examples in each mid-level moveme into multiple "fine-grained movemes", each corresponding to a specific human pose type (e.g. raise hand, reach, etc.). The intuition is that, though consistent in viewpoint, the mid-level movemes still cannot capture the level of details that are typically important for inferring the future actions, particularly when only a single frame or a short video clip is available. We propose to use fine-grained movemes to capture these human pose types (or atomic motions).

Fig. 4 shows an example of a mid-level moveme that contains two motion segments of persons with the same viewpoint, before the starting point of the action (high five). To avoid the confusion of terminology, we will use "motion segment" to denote the track associated with a person truncated at the starting point of the action we wish to predict, and "atomic motion segment" for the consecutive frames of a person which share a similar pose type, as shown in Fig. 4.

Our algorithm for discovering the fine-grained movemes consists of two steps: First, we cluster the frames in each person's motion segment independently. The goal is to find the atomic motion segments of each person which share a similar pose type. Second, we merge the different person's atomic motion segments that correspond to the same pose type into a fine-grained moveme. In this way, a fine-grained moveme contains multiple persons with consistent atomic motions (pose type). These two steps are explained in details below.

STEP 1. We develop a discriminative temporal clustering based method for finding the atomic motion segments for each person independently. Given all of the frames in a person's track prior to the action we want to predict, we cluster them based on appearance. These clusters will correspond to certain pose types. Every frame of the person is represented by a rigid HOG template. Instead of using the high-dimensional HOG representation for clustering, we train an exemplar SVM [13,20] for each person example, and use the detection score of each example to create a $K \times K$ similarity matrix. The (i, j) entry in the similarity matrix is the detection score of running the i-th detector on the j-th example. Once we have the similarity matrix, we cluster the frames of the person using a recently proposed temporal clustering algorithm [4]. We use a dynamic time warping (DTW) kernel

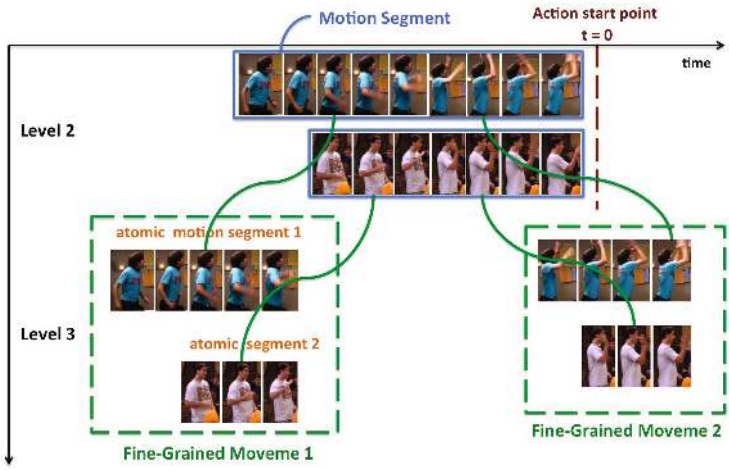


Fig. 4. Discovering fine-grained movemes. The figure illustrates how to discover the fine-grained movemes from the mid-level movemes. First, we cluster the frames in each person's motion segment to find the atomic motion segment of each person, which share a similar pose type. Then we merge the different person's atomic motion segments that correspond to the same pose type into a fine-grained moveme.

to achieve the invariance of temporal order, i.e. each cluster contains the atomic segment of the person with consecutive frames with the same order of the original sequence, as shown in Fig. 4.

STEP 2. The second step of our algorithm is to merge the atomic motion segments that correspond to the same pose type into a fine-grained moveme. For example, in Fig. 4, each of the discovered atomic motion segments correspond to a pose type of the human movement. Both of the atomic motion segments in the bottom left of Fig. 4 correspond to the first pose type, while the ones in the bottom right correspond to the second pose type. Atomic motion segments corresponding to the same pose type are merged into a fine-grained moveme. Thus each fine-grained moveme represents a particular pose type (e.g. raise hand, reach, etc.). We consider at most 3 pose types for each motion segment.

3.2 Learning a Collection of Moveme Classifiers

Given a hierarchy of movemes, we learn a classifier for each moveme in the hierarchy. Our goal is to predict future actions based on a single frame or a short video clip. Thus for each moveme, we learn two classifiers, based on appearance (HOG) and motion cues (HOF and MBH [22]), respectively. When the input is a single frame, we only consider classifiers trained with appearance features, while the input is a video clip, we consider both.

A coarse-level moveme models generic pose and viewpoint characteristics of certain action that is about to take place. Each motion segment within a moveme is associated to the same future action label. We compute feature descriptors for

persons at each frame and train a multi-class SVM on top of the feature representations. The learned SVM weights tells how likely the person will perform each action in the near future.

A mid-level moveme models viewpoint-specific but pose-generic characteristics of the future action. Each motion segment within a moveme is associated to the same viewpoint and future action label. For each moveme, we use all person bounding boxes that correspond to the moveme as positive examples, and random patches as negative examples. We then train a linear SVM for detecting the presence of the moveme.

A fine-grained moveme models viewpoint-specific and pose-specific characteristics of the future action. Each atomic motion segment within a moveme is associated with the fine-grained moveme label automatically discovered in the discriminative clustering process. We use the same strategy as defined above for training the fine-grained moveme classifiers. Examples of movemes and their corresponding templates are shown in Fig. 5.

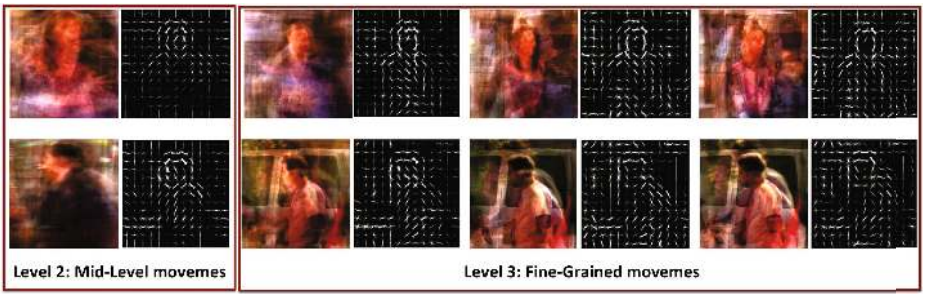


Fig. 5. Averaged images and moveme templates. We visualize the learned templates of the mid-level movemes and the fine-grained movemes in the hierarchy for hug (first row) and handshake (second row). For each template, we show the images averaged over all examples that belong to the same moveme.

4 Model

We introduce a model that is able to combine information across different movemes in a structured hierarchical way. It performs future action prediction and explicitly models the relations between movemes in different layers. Moreover, the model implicitly performs viewpoint prediction and temporal localization of the input frame (or short video clip) w.r.t. the start point of the action.

4.1 Model Formulation

The input to our learning module is a set of N video frames and short clips of persons. Each person example X is associated with labels corresponding to one

branch of the movemes hierarchy: $Y = \{y_i\}_{i=1}^{L}$, where L is the total number of levels of the hierarchy (we set it to 3) and y_i is the index of the corresponding moveme at level i. For example, y_1 corresponds to the future action label, y_2 corresponds to the label of a future action with a particular viewpoint (e.g. handshake while facing to left) and y_3 corresponds to the fine-grained moveme label that is automatically discovered by our clustering algorithm.

Our scoring function for labeling an example X with movemes Y is written as:

$$\Phi(X, Y) = \sum_{i=1}^{L} \alpha_{y_i}^{\top} \phi(X, y_i) + \sum_{i=1}^{L-1} \beta_{y_i, y_{i+1}}^{\top} \psi(y_i, y_j) \tag{1}$$

Unary model $\alpha_{y_i}^{\top} \phi(X, y_i)$: This potential function captures the compatibility between the example X and the moveme y_i. We use $\phi(X, y_i)$ to denote response of running the moveme classifier of y_i on the person example X. If X corresponds to a person track over a short clip, then we take the max response of the moveme classifier over all frames on the track. To learn biases between different movemes, we append a constant 1 to the end of each response.

Pairwise model $\beta_{y_i, y_{i+1}}^{\top} \psi(y_i, y_j)$: This potential function captures the compatibility between a pair of movemes located across different levels of the hierarchy. We write $\psi(y_i, y_j) = 1$ if the movemes y_i and y_j are connected by an edge in the hierarchy, and $-\infty$ otherwise. This means we exclude the co-occurrence of certain pairs of movemes: e.g. a person can not be described by movemes corresponds to the prior observation of different actions at the same time. Here $\beta_{y_i, y_{i+1}}$ is a model parameter that favors certain pair of movemes to be chosen for a person.

4.2 Inference

For an example X that corresponds to a person in a single frame or over a short video clip, our inference corresponds to solving the following optimization problem: $Y = \arg\max_{y_i : i=1,\ldots,L} \Phi(X, Y)$. For the example X, the inference is on a chain structure where we jointly infer moveme labels at all levels together. This is a simple exact inference and we solve it using Belief Propagation. The moveme at the top layer of the hierarchy y_1 corresponds to the future action label of the person. Our inference procedure also returns other more detailed predictions of the person (e.g. viewpoint, temporal state) through movemes at the other layers of the hierarchy (latent variables in our model) $\{y_i\}_{i=2}^{L}$.

4.3 Learning

Given a collection of training examples in the form of $\{X^n, Y^n\}_{n=1}^{N}$, we learn the model parameters θ that tend to correctly predict the future action labels. We formulate this as follows:

$$\begin{aligned} &\min_{\theta, \xi \geq 0} \ \frac{1}{2}||\theta||^2 + C \sum_n \xi_n \\ &\theta^{\top} \Phi(X^n, Y^n) - \theta^{\top} \Phi(X^n, Y^*) \geq \Delta\left(y_1^n, y_1^*, t\right) - \xi_n, \forall n, \end{aligned} \tag{2}$$

where $\Delta\left(y_1^n, y_1^*, t\right)$ is a loss function measuring the cost incurred by predicting y_1^* when the ground truth is y_1^n. Since our goal is to predict the future action labels, we only penalize the incorrect predictions of the future action label, rather than movemes in other layers of the hierarchy. A standard loss function of Structural SVM is the $0-1$ loss which equally penalizes all incorrect predictions at any time prior to the future action. However, this is inadequate for the task of future action prediction, since prediction from a frame at a long time before the start point of an action is obviously more difficult than from those at a few frames before the action is happening. If we treated them equally in training, then the learned decision boundaries might become unreliable.

Here we introduce a new loss function that depends on the temporal distance to the future action: $\Delta\left(y_1^n, y_1^*, t\right) = 1 - \mu t$ if $y_1 \neq y_i^*$, and 0 otherwise. If the example is in a sequence that does not contain any action of interest, we simply use the $0-1$ loss. Here $t \in (0, T]$ is the temporal distance to the starting point of the action we wish to predict, and $t = 0$ corresponds to the first frame of the action, T is the maximum number of frames before the action that we consider. $\mu \in (0, 1/T]$ is a tunable parameter. In this case, incorrect prediction from frames longer before the action is happening receives less penalties.

The optimization problem of Eq. 2 is convex and many well-tuned solvers can be applied to solve this problem. Here we use the bundle optimization solver in [2].

5 Experiments

Our goal is to test the performance of the proposed method on future action prediction in the challenging real world scenarios. At that end, we choose a very challenging dataset collected from TV shows [14], which include actions that we typically perform at a daily basis. We show that our method significantly outperforms baselines in future action prediction when the input is only a single frame or a short video clip.

The proposed method is generic and will not lose the discriminative power in classifying videos containing activities at relatively early stage or even the fully observed activities. We also evaluate our method on the UT-Interaction benchmark dataset [18]. We show that our method achieves state-of-the-art performance in early activity prediction.

Implementation Details. In all experiments, the penalty parameter C of the Structured SVM objective (Eq. 2) is set to 1 for both our method and the baselines. The codebook size for the dense trajectory descriptors [22] is set to 2000 for TV Interaction dataset and 800 for UT Interaction dataset.

5.1 TV Human Interaction Dataset

This dataset consists of 300 video clips collected from over 20 different TV shows. It contains five action classes: handshake, high five, hug, kiss and none. The class "none" represents all other more general actions such as walking and standing.

Annotations are provided for every frame of the videos, including the upper body bounding boxes, discrete head orientations and action labels for each person.

We use the training/testing split provided along with the dataset. For training, we sample a collection of frames and short clips from all of the videos in the training set, which contains more than $25,000$ person examples. This ensures that the system has "seen" a large number of videos on human actions before making a prediction. In testing, the experiments were conducted with different settings on the lengths of the input video clips as well as their temporal distances to the start point of the action we wish to predict (see below for details). In the most challenging scenario, we predict future actions from a static video frame.

Baselines. We compare our method agains the following baselines: 1) SIFT flow [12]. Given a testing image, it first finds the nearest neighbor from the training data using the SIFT flow algorithm, which matches densely sampled SIFT features between the two images, while preserving spatial discontinuities. The future action label of the matched training image is directly transferred to the testing image. 2) Dense flow [22]. We apply one of the state-of-the-art action recognition methods for future action prediction. The model is trained with video clips containing fully executed actions and tested for future action prediction. A linear SVM is used. 3) Our model with only the top most layer ("1-Layer"). 4) Our model with the top two layers ("2-Layer").

Results. We evaluate the performances when the input is a single image or a short video clip of four different lengths ($1, 3, 5, 7$ frames). All of the videos in this dataset have the same frame rate of 24 fps. Thus the longest video clip we provide at testing (7 frames) is less than 0.3 s, making the problem of future action prediction very challenging. Note that the input clip of length 1 denotes that we use a single frame as input, but with both shape and motion features.

We only use the shape feature (HOG) to represent the person when the input is a single image, and use both shape (HOG) and motion features (the dense trajectory descriptors [22]) when the input is a video clip. We set the trajectory length to 5 frames. Note that for each frame, the trajectories are computed using the feature points sampled from the five-frame temporal segment before the current frame. This guarantees that we don't have access to any *future* information in feature computation.

In order to test the methods' ability in predicting future actions at different stages, we measure the performances with 5 different temporal stage settings, from -20 to 0, with a step size of 5. The numbers denote the temporal distance (in frames) from the input image to the start point of the action. For example, the methods' performances at a temporal stage -20 describe the classification accuracies given all of the testing frames within 20 frames before the start point of the action we wish to predict. The temporal stage of 0 indicates all testing images are taken within 5 frames after the start point of the action, making the problem a conventional action classification problem. The comparative results are shown in Fig. 6. Our method outperforms all of the baselines at all different temporal settings. It is interesting to see that there is a notable performance increase of our

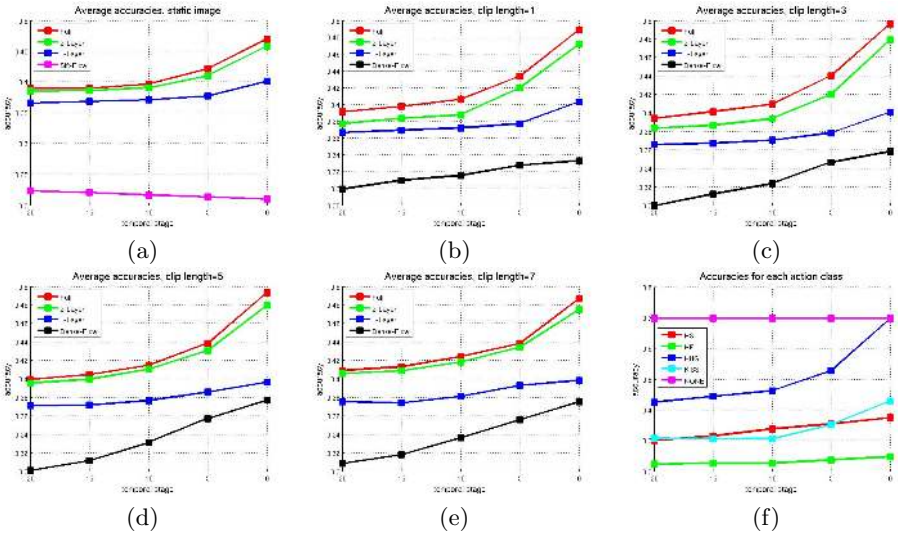


Fig. 6. Future action prediction accuracies. We evaluate performances when a static frame (a) or a short video clip (b)-(e) is available. The X axis corresponds to the temporal stages of the input frames, while the Y axis corresponds to the mean-per class accuracies. The red curve denotes our method, green for using the first two layers of the moveme hierarchy, blue for using the first layer, black for Dense flow [22] and magenta for SIFT flow [12]. (f) shows the accuracies for each action class given a single frame using the full model. HS and HF denote handshake and high five respectively.

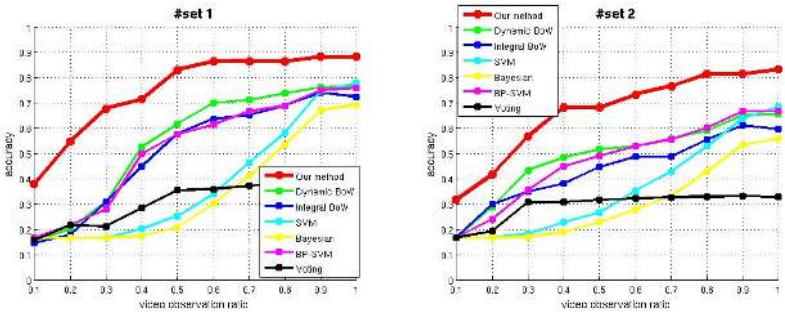
Fig. 7. Future action prediction visualizations. We show predictions of our method at different temporal stages before the action is executed (in yellow). For example, $t = -15$ in the first image denotes that the image is taken 15 frames before the action (handshake) starts. Correct predictions are shown in green and incorrect predictions in red. The last row shows examples of failure.

full model as well as the 2-Layer moveme model, starting from 10 (around 0.5 s) before the action is executed. This is because the fine-grained appearance and motion that characterize the actions tend to appear around 10 frames before the action starts. This can be verified by the visualization of our predictions shown in Fig. 7[1].

5.2 UT Interaction Dataset

The proposed hierarchical moveme representation is generic and captures the "multi-modality" nature of human movements. Thus its application domain is not limited to future action prediction, but also other aspects of human activity understanding, such as early action prediction and action recognition.

We validate the proposed method on the UT-Interaction benchmark dataset [17]. The dataset contains a total of 120 videos of 6 classes of human interactions (e.g. handshake, hug and kick). In order for a fair comparison with other reported numbers on this dataset, we follow the experiment settings as in [18]: we evaluate the proposed method on both Set #1 and Set #2 of UT-Interaction (segmented version), and use leave-one-sequence-out cross validation for measuring the performances. We run the person detector and tracker provided by [21] to obtain person tracks in the video sequences.



(a) (b)

Fig. 8. Early action prediction accuracies. The comparisons of the proposed model and all other methods (reported in [18]) on the UT-Interaction dataset.

Fig. 8 compares our model with existing methods on early activity prediction. Following [18], we tested our full model using 10 different observation ratios. Our method significantly outperforms all other methods on both # Set 1 and # Set 2 of UT-Interaction. Table 1 compares results of our method with leading approaches on the UT-Interaction dataset. Our method achieves state-of-the-art in terms of predicting activities at a relatively early stage, accessing only the first 50% of the testing video. An average classification accuracy of 83.1% is achieved, which is nearly 10% better than the current best result [16] on this benchmark.

[1] More visualizations are available at our website `http://cs.stanford.edu/$\sim$taranlan`.

Table 1. Performance comparisons on UT Interaction # 1. Table compares classification accuracies of our approach and the previous approaches; leading approach is in **bold**. The second and third columns report the accuracies using the first half and the entire video, respectively. Our method achieves state-of-the-art in recognition at an early stage when only the first half of the video is available, and outperforms the current best result [16] by nearly 10%.

Methods	Accuracy w. half videos	Accuracy w. full videos
Our model	**83.1%**	88.4%
Ryoo [18] (Avg.)	61.8%	76.7%
Ryoo [18] (Best)	70%	85%
Cuboid+SVMs [3] (Avg.)	25.3%	78%
Cuboid+SVMs [3] (Best)	31.7%	85%
BP+SVM [18] (Avg.)	57.7%	75.9%
BP+SVM [18] (Best)	65%	83.3%
Raptis & Sigal [16]	73.3%	93.3%
Yao et al. [26]	-	88%
Vahdat et al. [21]	-	93.3%
Zhang et al. [30]	-	**95%**
Kong et al. [8]	-	88.3

For action recognition, our number is slightly lower than state of the art [30], we think it is for two reasons: 1) our method is designed for prediction from a single frame or short video clip, so we don't model the temporal relations across frames over relatively long video sequences. 2) The use of linear versus complex kernels.

6 Conclusions

We have presented hierarchical movemes - a new representation for predicting future action from still images or short video clips in unconstrained data. Different movemes in our representation capture human movements at different levels of granularity. Movemes are organized in a structured hierarchical model and the model parameters are learned in a max-margin framework. Our experiments demonstrate that our model is effective in capturing the fine-grained details that are necessary for future action prediction. In addition, the model is generally applicable to other aspects of human activity understanding: the proposed model outperforms multiple state-of-the-art methods in early action prediction on a benchmark dataset.

Acknowledgments. We acknowledge the support from a Ford-Stanford Innovation Alliance Award and the ONR award N00014-13-1-0761.

References

1. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: CVPR (1997)
2. Do, T.M.T., Artieres, T.: Large margin training for hidden markov models with partially observed states. In: ICML (2009)

3. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: ICCV 2005 Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005)
4. Zhou, F., De la Torre, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. PAMI (2013)
5. Fanti, C.: Towards Automatic Discovery of Human Movemes. Ph.D. thesis, California Institute of Technology (2008)
6. Hoai, M., De la Torre, F.: Max-margin early event detectors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2012)
7. Kitani, K.M., Ziebart, B.D., Bagnell, D., Hebert, M.: Activity forecasting. In: European Conference on Computer Vision (2012)
8. Kong, Y., Jia, Y., Fu, Y.: Learning human interaction by interactive phrases. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 300–313. Springer, Heidelberg (2012)
9. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: Robotics: Science and Systems, RSS (2013)
10. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: Computer Vision and Pattern Recognition, CVPR (2012)
11. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2008)
12. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT flow: Dense correspondence across different scenes. In: European Conference on Computer Vision (2008)
13. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: IEEE International Conference on Computer Vision (2011)
14. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in tv shows. PAMI (2013)
15. Pellegrini, S., Ess, A., Schindler, K., Gool, L.J.V.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
16. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
17. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV (2009)
18. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: IEEE International Conference on Computer Vision (2011)
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: IEEE International Conference on Pattern Recognition, vol. 3, pp. 32–36 (2004)
20. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 73–86. Springer, Heidelberg (2012)
21. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: VS (2010)
22. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision (2013)
23. Wang, Y., Tran, D., Liao, Z., Forsyth, D.: Discriminative hierarchical part-based models for human parsing and action recognition. JMLR (2012)

24. Wang, Z., Deisenroth, M., Amor, H.B., Vogt, D., Scholkopf, B.: Probabilistic modeling of human movements for intention inference. In: Robotics: Science and Systems, RSS (2013)
25. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010)
26. Yao, A., Gall, J., Gool, L.V.: A hough transform-based voting framework for action recognition. In: CVPR (2010)
27. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010)
28. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: IEEE International Conference on Computer Vision (2011)
29. Yuen, J., Torralba, A.: A data-driven approach for event prediction. In: European Conference on Computer Vision (2010)
30. Zhang, Y., Liu, X., Chang, M.-C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 707–721. Springer, Heidelberg (2012)