Modeling Human Activities as Speech

Chia-Chih Chen and J. K. Aggarwal
Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
{ccchen | aggarwaljk}@mail.utexas.edu

Abstract

Human activity recognition and speech recognition appear to be two loosely related research areas. However, on a careful thought, there are several analogies between activity and speech signals with regard to the way they are generated, propagated, and perceived. In this paper, we propose a novel action representation, the action spectrogram*, which is inspired by a common spectrographic representation of speech. Different from sound spectrogram, an action spectrogram is a space-time-frequency representation which characterizes the short-time spectral properties of body parts' movements. While the essence of the speech signal is the variation of air pressure in time, our method models activities as the likelihood time series of action associated local interest patterns. This low-level process is realized by learning boosted window classifiers from spatially quantized spatio-temporal interest features. We have tested our algorithm on a variety of human activity datasets and achieved superior results.*

1. Introduction

Recognizing human activities from videos is one of the most intensively studied areas in computer vision. It is of significant interest in many applications, such as video surveillance, indexing, abnormal activity detection, eldercare, and human computer interactions. Compared to the research in automatic speech recognition (ASR), human activity recognition is a relatively young discipline. The first ASR system was built in the 1950s [9], and now the commercialized services and products are used in daily lives. These two seemingly unrelated areas share some very similar goals and processing methodologies. For example, we expect an ideal video surveillance system to accurately segment and semantically annotate continuous activities of multiple agents in unconstrained environments. Likewise, the ultimate goal of ASR is to segment and label spontaneous and continuous speech into constituent words then sentences independent of speakers and vocabulary. In addition, activity and speech are both temporal data; therefore, techniques such as Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) are commonly adopted for the recognition of activity and speech sequences.

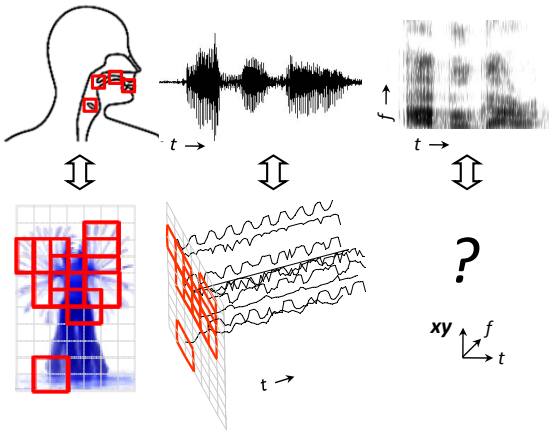


Figure 1. We compare human activities to speech, and introduce the analogies between articulatory apparatus and body parts, air pressure wave and local likelihood time series, and spectrogram and our spectrogram-like representation.

We are motivated to model human activities as speech due to the analogies between their production mechanisms. While speaking, part of our articulatory apparatus continuously reshape the vocal track which causes time varying resonances of the exhaled air flow. The magnitude of the propagated air pressure wave is a non-stationary signal which is relatively stationary when observed in short time intervals. Therefore, as shown in Fig. 1, one common way to characterize digitized speech signals is to extract the magnitude spectrum from each equally spaced and overlapped time window (frame in ASR). The representation that concatenates individual spectra in time is called a *spectrogram*. The time span of the analysis window is approximately equal to the period while the vocal track sustains its shape (10 to 50ms). This setup validates the assumption that each time segment of the speech signal is quasi-stationary [28].

On the other hand, the motion of human body parts

(limbs, head, torso, etc.) also emit time varying visual patterns at a relatively low frequency band. Nevertheless, if we are to compare body parts to articulatory apparatus, there are two minor differences to be clarified. First, it is mainly the shape of the articulatory apparatus that manipulate the articulation of phonemes, while human actions are distinguished by the simultaneous interest patterns (of motion or gesture) from parts. Second, for speech, the waveform of sound is already synthesized within the vocal track, while in action different body parts create different visual patterns and are perceived as a whole. In the ASR community, there has been an emerging interest in incorporating visual information for recognition. Lips are the most visible articulatory apparatus; therefore, various visual features [6] extracted from the corresponding area have been shown to further recognition performance. Despite all the "acting apparatus" being directly visible, there is little in the way of research exploring the temporal signals [1, 17] emitted from body parts for activity recognition (the sounds of actions).

Similar to speech signals, if we are able to model the associated interest patterns of an action at body part level, their occurrence likelihood in a short time period can be also deemed as quasi-stationary. Based on this observation, we propose a spectrogram-like representation to characterize human activities. We name it *action spectrogram* (AS). Compared to a 2D spectrogram, AS is a space-time-frequency representation which records the occurrence likelihood spectra of action specific interest patterns emitted from body parts. However, there are three major issues to be solved to make this kind of representation possible:

- how are local interest patterns defined and located?
- how are local interest patterns associated with actions?
- how do we model the occurrence time series of local interest patterns?

In this work, we provide a complete solution to these issues. First, we define local interest patterns as the video content indicated by the spatio-temporal interest points (STIP) [10] within a figure-centric action sequence. Second, to associate local interest patterns with actions, we modify Adaboost algorithm to learn a set of action associated spatio-temporal interest point detectors (AASTID) from each action. Third, we use the boosted AASTID to compute the occurrence likelihood of local interest patterns from different body parts. These likelihood time series are divided into overlapped short time segments (likelihood segments) and converted by an 1D Fast Fourier Transformation (FFT) to synthesize AS. We train Support Vector Machines (SVMs) to classify an activity AS into the component actions.

Our work provides a novel perspective to the characterization of human activities, which may induce the transfer of research in both areas of speech and activity recognition. We not only make the associations between different aspects of speech and activity signals, but also contribute a viable solution to recognize continuous activities as speech. The remainder of the paper is organized as follows. §2 summarizes the related work. §3 introduces the technical details of AS computation. In §4, we present the methodologies to classify a single AS slice and a continuous AS sequence as a stream of activity. We demonstrate our experimental results on 4 diverse datasets in §5, and conclude in §6.

2. Related Work

Our spectrogram-like representation of activities is a type of mid-level feature [12, 10, 18, 21], which is built upon low-level features such as image gradients or optical flow. Compared to features that describe the entire human figures, mid-level features are focused on local regions of an action sequence to provide efficient yet descriptive representation.

As shown in the middle of Fig. 1, speech signals are compared to the occurrence likelihood of local interest patterns in time. These likelihood time series appear to be very similar to the trajectories of body parts; however, they are essentially different. For example, Matikainen *et al.* [17] employ a feature tracker to track a number of features over an activity sequence. The trajectories of the tracked features are processed and divided into snippets called *trajectons*. Under a bag-of-words framework, tracjectons of a video are matched against a pre-clustered trajecton library and accumulated into a histogram-based action descriptor. Different from [17], Ali *et al.* [1] assume that feature tracking is a relatively well-solved problem so that the trajectories of body reference joints can be reliably used as a feature. The focus of their method is to model the nonlinear dynamics of human actions by the theory of chaotic systems.

The use of local spatio-temporal video features to represent activities has drawn considerable interest in the past few years [23, 15, 16, 18, 21]. The work by Wang *et al.* [26] provides a comprehensive performance evaluation on different combinations of popular local spatio-temporal feature detectors and descriptors. Ke *et al.* [15] and Laptev and Pérez [16] boost a cascade of space-time window classifiers to recognize actions. To make the runtime scalable, their weak learners are trained on features extracted from random cuboids of dense video grids. Neibles *et al.* [18] represent an image sequence as a bag of video words. Under their unsupervised learning framework, action recognition and localization are performed by maximizing the posterior of learned category models. Ryoo and Aggarwal [21] propose a kernel function to measure the structural similarity between the sets of STIP extracted from two videos. Their kernel is a histogram which bins the pairwise spatio-temporal relationships among the video words.

In previous work on spectral analysis of human action, Cutler and Davis [7] detect periodic motion by analyzing

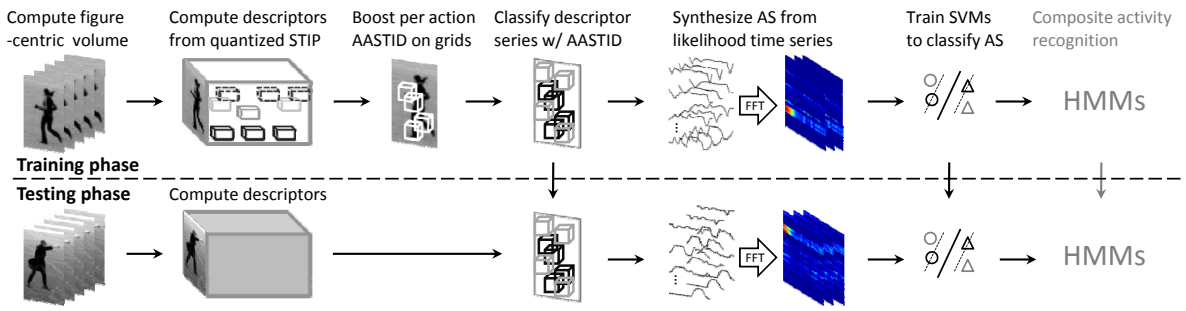


Figure 2. Flow diagram of our activity recognition scheme. The vertical arrows indicate the supply of trained models.

the power spectrum of the sequence self-similarity matrix. Weinland *et al.* [27] propose a free viewpoint action descriptor based on Fourier analysis of motion history volumes. The shift invariant property of FFT enables them to extract view-invariant features from cylindrical coordinates.

3. Action Spectrogram

An important process in the computation of AS is to quantize the occurrence time series of action specific local interest patterns. This involves the stabilization of human figures, the learning of action associated local patterns, and the estimation of occurrence likelihood. Also, we need to evaluate the proper time interval to divide continuous likelihood series into short segments, which are synthesized into AS. This process is a part of the overall algorithm as in Fig. 2.

3.1. Preprocessing and Action Features

Preprocessing. Our algorithm, similar to [11, 12], operates on the figure-centric spatio-temporal volume of activity. Depending on the setup of the activity recognition system, this generally requires detection and continuous tracking of human objects. In most of our tested datasets [2, 23, 22], there is one human object in a frame; therefore, we perform tracking by human detection.

Given a raw video stream, the goal of our preprocessing is to extract the sequence of figure-centric bounding boxes. We adopt an approach similar to [8] for human figure stabilization. Our human figure stabilizer is a linear SVM classifier which is trained with Histograms of Oriented Gradients (HOG) [8] descriptors extracted from manually cropped figure-centric bounding boxes and negative examples from random patches around the figures. We perform a multi-scale human detection in the local neighborhood defined by the previously detected location. We keep track of the stabilized bounding box coordinates and the corresponding HOG vectors for later processing.

Action Features. We use both shape and motion histogram based features to characterize human activities. In addition to the performance benefits, combining features of different types provides a broader coverage of activities. For example, there are scarce features due to motion which can be extracted to distinguish certain static actions such as 'stand' from 'gesture'. More specifically, within a figure-centric volume, we represent successive poses by time series of HOG and motions by time series of Histogram of Oriented Optical Flow (HOF) [16].

HOG descriptor divides the subject figure into equally spaced regions called cells, and represents the edge structure of each cell by the angular distribution of gradients. Compared to a cell, a block covers a larger region which consists of several cells. In [8] the cell histograms within a block are normalized to provide better invariance to illumination and shading. Here we characterize the appearance of human body parts at the spatial scale of a block. The overlapped blocks are more robust against minor stabilization errors and describe parts with the context of adjacent cells. As shown in Fig. 3(a), our implementation uses 2×2 cell blocks and follows the common settings as in [8]. Note that we compute HOG time series via figure stabilization.

We describe the motion field between each pair of successive figure-centric frames by HOF descriptor. Besides the types of feature being characterized, the main difference between HOG and our modified HOF descriptor is the orientation mapping carried out. We follow [8] to use unsigned gradient vectors in HOG computation. In general applications, the acting directions of a person are not used as a cue to distinguish actions. Therefore, to make optical flow vectors symmetric about the vertical axis, the orientation of a flow vector is converted by

$$\theta_{of} = \begin{cases} sgn(\theta_{of}) \cdot \pi - \theta_{of}, & |\theta_{of}| > \frac{\pi}{2} \\ \theta_{of}, \quad otherwise. \end{cases} \tag{1}$$

The cells of the HOF descriptor capture the relative motions of parts at a finer spatial scale. The same as HOG, we describe the motion patterns of parts in every 2×2 cell block.

3.2. Learning Action Associated Interest Patterns

We compare the learning of action associated local interest patterns to the search of correspondence between a uttered phoneme and the shapes of those active articulatory

apparatus. Actions appear to vary across both time and body parts; however, not every local video feature contributes to the correct recognition of actions. One effective technique to select a variety of discriminant features is to evaluate the weak classifiers trained on an overcomplete set of features [24]. Our method is similar to Ke *et al.* [15] and Laptev and Pérez [16] in the sense of discovering discriminant cubic features in a boosting framework. Nevertheless, our work differs from theirs with regard to the method of selecting boosting instances and the format of action classifiers as suggested in §2 and to be detailed later.

In general, STIP detectors are used to localize local video structures which pose significant variations in both space and time. Our AASTID are boosted space-time window classifiers, which are *not* trained to detect points of interest but to produce the occurrence likelihood of action specific STIP. Here we assume, within the figure-centric volumes of the same action, the STIP that are in close *spatial* proximity of each other present similar interest patterns. One important observation that motivates us to boost AASTID from STIP is that action associated interest patterns occur in an intermittent fashion. For example, in a spatio-temporal volume of a person 'kicking', the most descriptive video cuboids cover the sweeping leg in time. However, after the leg goes down, there is no subsequent interest pattern emitted from the leg position until the next kick. Previous methods such as [15, 16] select boosting examples by randomly sampling cubiods from dense video grids. Their approach inevitably includes positive features from video cuboids which do not relate to the action (*e.g.* arbitrary background volumes) and negative features which do not characterize the rest of the actions well. As a result, the discriminating power of the boosted weak classifiers are weakened by labeling uninformative video cuboids as positive and negative examples (Fig. 3(b) for example).

We use the occurrence likelihood series of action associated STIP as features. Ideally the likelihood signal emitted from an AASTID is expected to peak, bottom, and level (about 0.5) when classifying features from positive, negative, and random video cuboids, respectively. We detail the implementation of AASTID as follows.

Extracting STIP. The popular STIP detector proposed by Dollár *et al.* [10] is a combination of a 2D Gaussian spatial kernel and 1D Gabor temporal filters. Their STIP detector is devised to be responsive not only to periodic motions but also to a wide range of other interesting space-time patterns. Via their implantation, we are able to extract a dense set of STIP to capture the details of a training volume. As shown in Fig. 3(a), a STIP response volume is computed using $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$, where $g(x, y; \sigma)$ is a Gaussian smoothing kernel and $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ are a quadrature pair of Gabor temporal filters. STIP are fired at local maxima by applying non-maximum suppression on the R volume. We quantize a local maxima to the grid location of a 2×2 cell block while maintaining its temporal location. This is achieved by comparing the integrals of R within the quantized video cuboids (section of a block) which overlap with the local maxima. We compute time series of HOG and HOF features from the quantized video cuboid with the maximum R integral. We denote a STIP of action α by $cb_\alpha(u, v, t)$, which is characterized by $\mathbf{h}(u, v, t, \theta)$ vectors, where (u, v) is the quantized grid location, t represent the time and the corresponding training volume, and θ indicates the type of histogram feature.

Boosting AASTID. We boost a set of AASTID per action. These detectors are mostly localized at the related body parts (see Fig. 1). Unlike [16], for reliable estimation of STIP occurrence likelihood, we employ instance weighted linear kernel SVM [3] for weak learners

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} C_i\xi_i \quad \text{subject to } \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0. \tag{2}$$

In this primal problem, the inverse of margin width together with the weighted sum of training errors are being minimized. C_i and ξ_i correspond to the penalty and training error of the instance-label pair $(\mathbf{x}_i, y_i)$. This SVM formulation enables a weak learner to minimize the classification error of samples weighted by previous boosting iterations. Our weighted SVM based weak classifiers are more robust than those of weighted Fisher Linear Discriminant [16] based given the limited number of training instances.

We modify Adaboost to learn AASTID from spatio-temporally scattered STIP. We follow the basic settings as in [24] and focus on presenting the differences. The set of AASTID boosted from action α are among the best weak learners D_α of the total $(nr - 1) \times (nc - 1) \times nf$ weak learners d_α, where nr, nc, and nf are the numbers of cell rows, columns, and feature types. For each grid location (u, v), we denote the set of all STIP time instances as $T(u, v)$. A weak classifier $d_\alpha(u, v, \theta)$ is learned to distinguish θ represented $cb_\alpha(u, v, T_\alpha)$ from $cb_{\neg\alpha}(u, v, T_{\neg\alpha})$, where $T_\alpha \cup T_{\neg\alpha} = T(u, v)$ and $T_\alpha \cap T_{\neg\alpha} = \emptyset$. The weighting of $\mathbf{h}(u, v, t, \theta)$ at iteration i is $w_i(u, v, t, \theta)$, which is updated by intersecting t with $T(u_{best}, v_{best})$ of iteration $i-1$ $(i > 1)$. $w_i(u, v, t, \theta)$ is updated to $\frac{\epsilon_{i-1}}{1-\epsilon_{i-1}} w_{i-1}(u, v, t, \theta)$ if and only if $cb(u, v, t)$ is only temporally overlapped with the correctly classified $cb(u_{best}, v_{best}, T_{correct})$ where $T_{correct} \subset T(u_{best}, v_{best})$ and ϵ_{i-1} is the minimum weighted error in $i - 1$. This implies that any $cb(u, v, t)$ overlap with the wrongly detected $cb(u_{best}, v_{best}, T_{wrong})$ or missing a temporal intersection will be emphasized in the next iteration.

Estimating likelihood. Similar to the preprocess of speech signals, our weak learners are trained to output calibrated

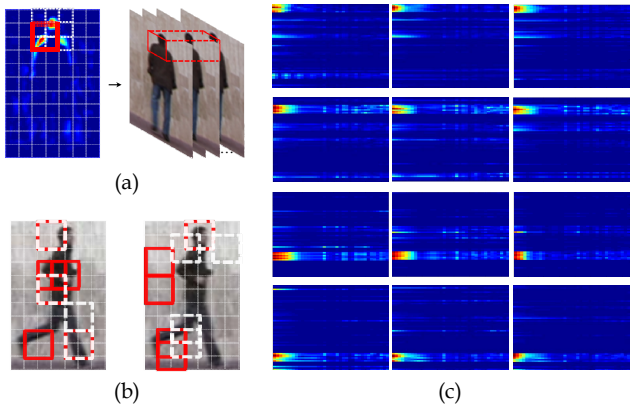


Figure 3. (a) Left: a slice of a STIP response volume. By referring to it, we quantize a local maximum at the head position to a grid location. (b) Left: D_{run} boosted from quantized STIP as in (a). Right: D_{run} boosted from dense video grids. The solid squares are gradient based D_{run}, and the dashed ones are optical flow based. The D_α computed by our method effectively capture the action associated body parts instead of some random background. (c) The sample AS time slices from the sequences (columns) of different actions (bend, jack, walk, wave1 in row) from [2].

likelihood values. Given the histogram vector, $\mathbf{h} \in R^n$, and the indicator of α, $y \in \{0, 1\}$, we aim to estimate the posterior probability $p(y = 1|\mathbf{h})$ using D_α. The method proposed by Wu *et al.* [29] approximates the posterior probabilities by coupling them with pairwise class probabilities. They start with modeling each pairwise class probability as a sigmoid of the corresponding decision value f

$$p(y = i|y = i \cup j, \mathbf{h}) \approx \frac{1}{e^{Af+B}}, i \neq j \tag{3}$$

where A and B are obtained by minimizing the negative log-likelihood function while f is calculated by performing cross-validation on the training set. The formulation of their pairwise coupling is based on the Bayesian equality

$$\begin{aligned} &p(y = i|y = i \cup j, \mathbf{h})p(y = j|\mathbf{h}) \\ &\quad = p(y = j|y = i \cup j, \mathbf{h})p(y = i|\mathbf{h}). \end{aligned} \tag{4}$$

This equality simply suggests that $p(y = i|\mathbf{h})$ is proportional to $p(y = i|y = i \cup j, \mathbf{h})$ in a binary problem, while it requires convex optimization for a multi-class problem.

3.3. Synthesizing Action Spectrogram

Compared to a sound spectrogram, the additional dimension of space in our representation characterizes the spatially distributed AASTID. We classify a figure-centric action volume with the spatial array of AASTID and synthesize the i^{th} AS slice from the frame interval $< (i-1)l_{step}+1, (i-1)l_{step} + l_D - 1 + l_{seg} >$, where l_{step}, l_D, l_{seg} are the temporal lengths of sampling step, AASTID, and likelihood segment, respectively. From each $l_D - 1 + l_{seg}$ frame sampled snippet of the volume, we can extract n_D length l_{seg} likelihood segments, where n_D is the total number of AASTID. The likelihood segments of a snippet are transformed by FFT and concatenated along the dimension of space to form a 2D time slice of the AS volume. We show the sample AS slices from 12 sequences of 4 actions in Fig. 3(c), where one action is distinguished not only by the active AASTID responses (bright rows) but also by its *spectral signature* (bright columns). For effective characterization of action, the selection of AASTID and the estimation of l_{seg} are introduced.

Selecting AASTID. As we boost the best weak learners on the spatial grids, they represent the most valid weak hypotheses about the action in the measure of detection rate; however, it is their spectral waveforms that are directly used as features. Therefore, we trim the best weak classifiers of each action to form the contributed set of AASTID. Let $D_\alpha(i)$ be the i^{th} best weak classifier of α, where i represents the trippet of (u_i, v_i, θ_i). We classify both the positive and negative ($\neg\alpha$) training volumes with $D_\alpha(i)$ and divide the emitted likelihood time series into n^+ and n^- fixed length segments. The spectra of the segments are denoted as $\{\mathbf{x}^+_{1,i}, \mathbf{x}^+_{2,i}, ..., \mathbf{x}^+_{n^+,i}\}$ and $\{\mathbf{x}^-_{1,i}, \mathbf{x}^-_{2,i}, ..., \mathbf{x}^-_{n^-,i}\}$. The discriminative value, $F(i)$, of $D_\alpha(i)$ emitted spectra is formulated as a Fisher discriminant like score [5]

$$\frac{\left\|\bar{\mathbf{x}}^+_i - \bar{\mathbf{x}}_i\right\| + \left\|\bar{\mathbf{x}}^-_i - \bar{\mathbf{x}}_i\right\|}{\frac{1}{n^+-1}\sum_{j=1}^{n^+}\left\|\mathbf{x}^+_{j,i} - \bar{\mathbf{x}}^+_i\right\| + \frac{1}{n^--1}\sum_{j=1}^{n^-}\left\|\mathbf{x}^-_{j,i} - \bar{\mathbf{x}}^-_i\right\|} \tag{5}$$

where $\bar{\mathbf{x}}_i, \bar{\mathbf{x}}^+_i$, and $\bar{\mathbf{x}}^-_i$ are the mean spectra of the entire, positive, negative training sets. The D_α with top F values are selected as the contributed set of AASTID from α.

Estimating l_{seg}. One popular approach to analyzing activities is to divide a video into snippets of frames and perform recognition from the snippets. In most of the literature, the duration of individual snippets is decided heuristically. Our speech-like representation of action provides a ready medium to tackle this problem. That is, by assuming each action is a random process, we can approximate the proper l_{seg} by performing a stationarity test on its realizations (likelihood series). Common methods for the test of stationarity include auto-correlation function and runs test [14]. They all require a sufficient number of samples per realization to make a meaningful judgement; however, most of the dataset videos are shorter than 3 seconds and sampled at a relatively low frame rate. Besides, these tests do not provide a normalized measure to indicate the degree stationarity.

We propose to approximate l_{seg} by calculating the average pairwise spectral similarities over segment lengths. In Fig. 4, as we sample longer and longer likelihood segments from the same action of [2], the corresponding AS slices

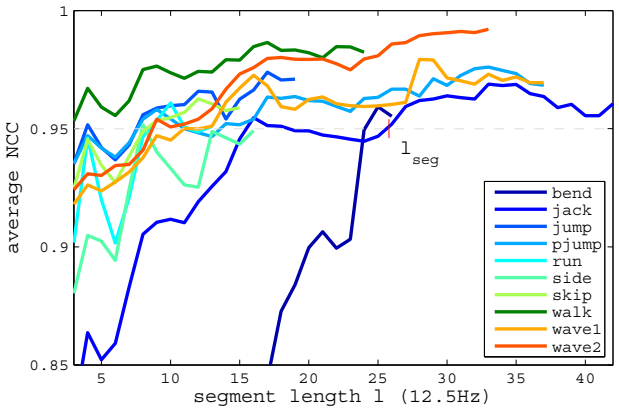


Figure 4. The average spectral similarities of AS as functions of l, which are used to determine l_{seg}. The likelihood segments are sampled with less than $\frac{1}{2}$ temporal overlap. The length of a curve depends on the duration of its longest sequence.

converge gradually in waveforms. The average pairwise similarity of the n AS slices of α is computed by

$$S(l,\alpha) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j:j\neq i} NCC(\mathbf{X}_i^\alpha, \mathbf{X}_j^\alpha) \quad (6)$$

where $(\mathbf{X}_i^\alpha, \mathbf{X}_j^\alpha)$ represents a pair of flattened 1D AS slices synthesized from length l segments, and NCC is short for Normalized Cross-Correlation. Given the target correlation value, we approximate a sufficient segment length, l_{seg}, by thresholding the similarity curves. The AS of aperiodic actions such as 'bend' require longer l to capture the complete occurrence. Note that we can certainly use a large l_{seg} to meet the target correlation value; however, this inevitably reduces the time resolution of the recognized activity.

4. Classification

We train a collection of one-against-one linear SVM classifiers [3] to recognize the AS slices of different actions. We prefer linear SVMs to other linear classifiers because they are rather discriminant while providing better out-of-sample generalization. Moreover, compared to nonlinear classifiers, they are easy to train, fast to run, and achieve consistently decent performance on different datasets and feature settings. We have tested several nonlinear kernel SVMs on our spectral data, for example, RBF, multi-channel Gaussian [30], and NCC kernel [25]. In our experiments, these nonlinear SVMs usually perform similarly or slightly better than the linear ones; however, their testing accuracies are sometimes subject to overfitting.

To recognize composite human activities, we consider a hybrid HMM approach [13], which has been implemented for real-world ASR applications. Traditional HMM based ASR systems model the state emission probabilities of phonemes using mixtures of Gaussians, which are replaced by more sophisticated classifiers such as Artificial Neural Networks (ANN) or SVM in a hybrid system. With the states corresponding to phonemes, spoken words are modeled by individual HMMs. Likewise, we slice-wise classify an activity AS into a sequence of actions, and model the temporal evolution of the sequential actions via activity HMMs. Our linear SVMs are trained to output the posterior probabilities (see §3.2) of actions, which can be applied to activity HMMs. Note that using our representation, interaction types of activities can be modeled by specialized HMM; for example, Oliver *et al.*[20] use coupled HMM to recognize person-person interactions.

5. Experimental Results

Fig. 5 summarizes the 4 datasets adopted to evaluate our method. The challenges posed by these datasets include low-resolution, blurry imagery, shadows, broken tracks, and variations in viewpoints, scales, scenes, lighting conditions, and clothing. We follow the same principles to initialize the parameters across datasets. For the computation of orientation histograms, we use 9 bin histograms, and set the block size approximately equal to $\frac{2}{3}$ of the limb length with the stride (block overlap) of a cell size. The histogram time series extracted from a video cuboid is normalized with L2-Hys [8]. We manipulate the values of σ and τ so that there are about 20 to 50 STIP fired per second depending on the complexity of the training action. Based on Eq. 5, no more than 10 AASTID per action are selected among the best weak learners which score less than a 45% error rate. To speedup runtime, we reduce the video frame rate to half of the original, but double the time resolution of likelihood series by spline interpolation. $2^{\lceil \log_2^{l_{seg}} \rceil}$-point FFT is performed to synthesize AS. For videos shorter than l_{seg}, we replicate the existing likelihood series so that there is at least one AS slice formed per sequence. Our experimental results and findings are detailed as follows.

Weizmann. The Weizmann dataset [2] was filmed at medium resolution in a controlled environment. This dataset consists of 93 sequences of 10 actions performed by 9 individuals. We apply our preprocessing technique to extract figure-centric volumes, because some of the provided foreground masks contain incomplete figures (e.g. shahar_side). Evaluated with leave-one-sequence-out cross-validation (LOOCV), our method achieves 100% accuracy on this dataset.

KTH. Similar to the resolution setting of Weizmann, KTH [23] is a much more challenging dataset. As shown in Fig. 5(b), KTH is comprised of 6 actions, which were taken at varied scales with persons wearing different clothing in different scenes. The entire dataset contains 2,391 short clips acted by 25 individuals. We follow the setup as in [23] to partition the dataset into 3 parts by person identity. We

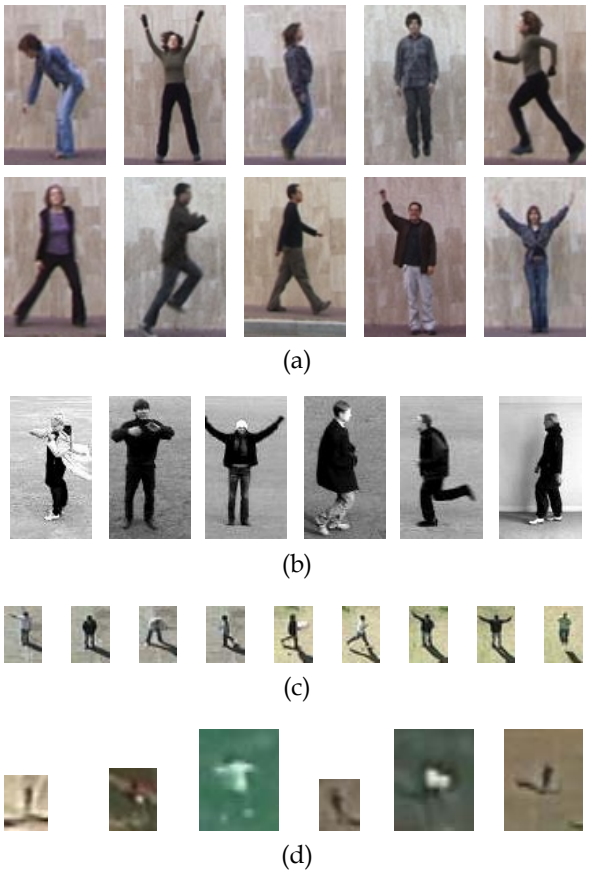
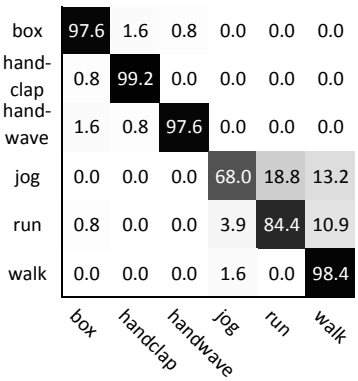


Figure 5. We tested our method on 4 datasets: (a) Weizmann (b) KTH (c) UT-Tower (d) VIRAT Aerial Video. The actions are self-explanatory from the figures except those from the Aerial dataset, where the actions are 'stand','dig','throw','walk','carry', and 'run'.

use $\frac{2}{3}$ of the dataset for training and the other $\frac{1}{3}$ for testing. Our linear SVMs correctly recognize 94.4% of the AS slices in the testing set. The average accuracy per action is 90.9%. The confusion matrix together with the comparison with other reported methods are tabulated in Table 1.

We are surprised to find that the per-video accuracy is about 3.6% lower than the per AS slice accuracy (90.8% *v.s.* 94.4%). After examining the error sequences, it is discovered that a significant portion of the misclassified clips are shorter than l_{seg} (1.5 seconds); however, these short clips represent 27% of the test set. Therefore, we conjecture that the disturbing likelihood spectra caused by an insufficient number of samples ($< l_{seg}$) and padding artifacts have led to the high error rate in short clips.

UT-Tower. The UT-Tower dataset [4] is a low-resolution dataset where actions were filmed top-down in a near aerial view and the human figures are 20 to 30 pixels in height. This dataset is composed of 9 actions performed by 6 persons in 2 scenes. Each subject repeats the same action twice



Method	ACC %
Ryoo [21]	91.1
Proposed	**90.9**
Fathi [12]	90.5
Niebles [18]	83.3
Dollár [10]	80.2
Schuldt [23]	71.7
Ke [15]	63.0

Table 1. Our results on the KTH dataset: the confusion matrix for per-video classification and the comparison with other methods.

so that there are 108 sequences. We perform LOOCV to compare with other methods as in [22]. The accuracy of our method is 98.2%, which is the best result reported on this dataset so far. The two incorrectly classified sequences are the 9^{th} sequence of 'walk' and the 5^{th} sequence of 'wave2', in which the low color contrast between a person's clothes and background confuses the classifier.

VIRAT Aerial Video. For the previous 3 datasets, our speech-like representation and recognition strategy demonstrate results that are better than or comparable to the state-of-the-art. To test the effectiveness of our methodology, we challenge it with video sequences taken from a Unmanned Aerial Vehicle (UAV). We manually select 42 sequences out of 6 actions from a large collection of UAV recorded footage named the VIRAT Aerial Video dataset [19]. The resolution of the videos is 720×480 pixels with the tracks of objects computed at 10 fps. As shown in Fig. 5(d), the imagery taken from an UAV not only creates difficulties due to low figure resolution, but also poses problems with vague object appearances, salient shadows, interrupted tracking (person temporarily out of FOV), and time varying viewpoints and scales. Due to these issues, part of the footage even requires repeated human scrutiny to perform ground truth annotation. Therefore, to propose a meaningful evaluation set, we select tracks of human actions which do not require a second inspection for labeling.

We refine the tracks with the preprocessing step to acquire stabilized action sequences. Even with this additional process, the quality of the extracted bounding boxes cannot be as consistent as those acquired from the other 3 datasets (see Fig. 5). To assess the performance of our method, we compare our accuracy with that of a baseline approach. We adopt time series of HOG extracted from overlapped spatio-temporal volumes (match the AS computation intervals) as the baseline descriptor. For the sake of fair comparison, we train linear SVMs on the HOG descriptors and use LOOCV accuracy as a measure. The average accuracy of our method is 38.3%, while it is 33.3% for the baseline approach. The

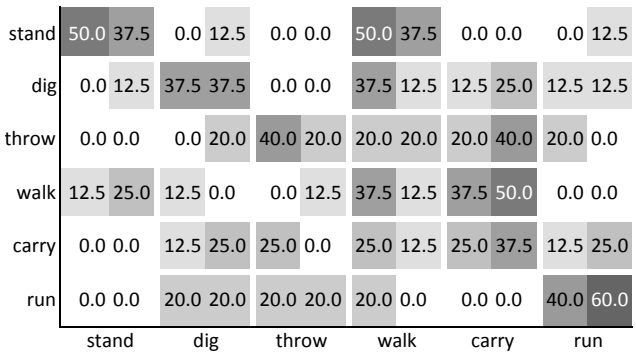


Table 2. The confusion matrices of ours (AS) and a baseline method (HOG time series) on the selected VIRAT Aerial Video dataset. The pair of percentages in each bi-colored cell represent our/baseline accuracy. The overall accuracies are 38.3% *v.s.* 33.3%.

confusion matrices are summarized in Table 2.

6. Conclusion

We have presented a novel activity recognition scheme which adapts naturally from ASR. We use both local video content and occurrence likelihood spectra to verify actions. More specifically, localized at body parts, the AASTID are trained to be responsive only to action specific interest patterns. The proposed AS is used to describe the temporal evolution of the ASSTID emitted likelihood spectra. The speech-like representation and recognition scheme offer two major advantages. First, we transform an activity sequence into simultaneous temporal signals, which enable us to analyze activities with signal processing techniques (*e.g.* §3.3). Second, we model activities as the composition of speech, which facilitates the evaluation of higher level activities with linguistic-like models. Our method demonstrates the feasibility of representing human activities as speech-like signals, which enables the further analysis of activities by various state-of-the-art speech recognition technologies.

7. Acknowledgement

This material is based upon the work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0135.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. *In ICCV*, 2007.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *In ICCV*, 2005.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] C.-C. Chen and J. K. Aggarwal. Recognizing human action from a far field of view. *In WVMC*, 2009.

[5] Y.-W. Chen and C.-J. Lin. *Combining SVMs with various feature selection strategies*. Springer, 2006.

[6] C. C. Chibelushi, F. Deravi, and J. S. D. Mason. A review of speech-based bimodal recognition. In *IEEE Trans. Multimedia*, volume 4, pages 23 –37, 2002.

[7] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. In *PAMI*, volume 22, pages 781–796, 1999.

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

[9] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. In *J. Acoust. Soc. Am.*, 1952.

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

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

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

[13] A. Ganapathiraju, J. E. Hamaker, and J. Picone. Applications of support vector machines to speech recognition. In *IEEE Trans. Sig. Proc.*, volume 52, pages 2348 – 2355, 2004.

[14] J. D. Gibbons. *Nonparametric Methods for Quantitative Analysis*. American Sciences Press, 1985.

[15] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. *In ICCV*, 2005.

[16] I. Laptev and P. Pérez. Retrieving actions in movies. *In ICCV*, 2007.

[17] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. *In ICCV Workshops*, 2009.

[18] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *IJCV*, volume 79, pages 299–318, 2008.

[19] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. *In CVPR*, 2011.

[20] N. M. Oliver, B. Rosario, and A. P. Pentland. A bayesian computer vision system for modeling human interactions. In *PAMI*, volume 22, pages 831–843, 2000.

[21] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *In ICCV*, 2009.

[22] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury. An overview of contest on semantic description of human activities 2010. *In ICPR Contests*, 2010.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. *In ICPR*, 2004.

[24] P. Viola and M. Jones. Robust real-time face detection. In *IJCV*, volume 57, pages 137–154, 2004.

[25] C. Wallraven. Recognition with local features: the kernel recipe. *In ICCV*, 2003.

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

[27] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. In *CVIU*, 2006.

[28] M. Woelfel and J. McDonough. *Distant Speech Recognition*. Wiley, 2009.

[29] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. In *JMLR*, volume 5, pages 975–1005, 2004.

[30] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. In *IJCV*, volume 73, 2007.