Segmenting and Recognizing Human Action using Low-level Video Features

Daphna Buchsbaum, Kevin R. Canini, Thomas L. Griffiths
daphnab@berkeley.edu, kevin@cs.berkeley.edu, tom_griffiths@berkeley.edu
University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Dividing observed human behavior into individual, meaningful actions is a critical task for both human learners and computer vision systems. An important question is how much action structure and segmentation information is available in the observed surface level motion and image changes, without any knowledge of human pose or behavior. Here we present a novel approach to jointly segmenting and recognizing videos of human action sequences, using a hierarchical topic model. Video sequences are represented as bags of video words, automatically discovered from local space-time interest points. Our model jointly infers both action identification and action segmentation. Our results are a good fit with human segmentation judgments as well as providing relatively accurate action recognition and localization within the videos.

Introduction

When observing a continuous stream of human behavior, an ongoing problem for both human learners and computer vision algorithms is recognizing and segmenting out individual, meaningful actions from within the longer motion sequence. This ability to pick out individual actions or events is thought to play an important role in human social cognition (Baldwin & Baird, 2001), allowing us to recognize other's behavior, and helping us link people's actions with their internal mental states, and with the outcomes of those actions (Buchsbaum, Griffiths, Gopnik, & Baldwin, 2009; Zacks, Speer, Swallow, & Maley, 2010). It is also a critical first step in developing computational approaches that can recognize and understand human action from naturalistic videos. Prior research has shown that adults are able to segment videos of common everyday activities into coherent actions, consistent with the goals and intentions underlying the actor's behavior e.g. (Newtson, Engquist, & Bois, 1977; Zacks, Tversky, & Iyer, 2001), and that even young infants are sensitive to the boundaries between intentional action segments (Baldwin, Baird, Saylor, & Clark, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007). People's boundary judgments are also sensitive to the hierarchical structure of goals and sub-goals underlying human action – they are able to consistently segment actions at multiple levels of granularity (Zacks, Tversky, & Iyer, 2001; Zacks, Braver, et al., 2001; Hard, Tversky, & Lang, 2006).

While a full understanding of human action requires knowledge about goals and intentions, infants are able to parse dynamic human action well before they are thought to have a fully developed theory of mind. This suggests that there may also be low-level cues to intentional action structure available in human motion, an idea supported by a variety of recent work (Zacks, 2004; Hard et al., 2006; Zacks, Kumar, Abrams, & Mehta, 2009; Meyer, DeCamp, Hard, & Baldwin, 2010; Hard, Recchia, & Tversky, under review). In particular, movement features within simple animated scenes (Zacks, 2004), as well as movement and position of tracked body parts within videos of everyday action (Zacks et al., 2009; Meyer et al., 2010), have been shown to correlate with human segmentation judgments. However, this previous work still assumed some a priori knowledge of human body pose, and of the types of motion features relevant to boundary detection. In this paper, we present a computational model that makes very few representational assumptions about what is observed, in order to explore the amount of action structure that can be inferred from just low-level changes in pixel values, without knowledge of human body structure, higher level goals and intentions, or even foreground/background distinctions. To the extent that the model corresponds to human segmentation judgments, and correctly recognizes actions, we know that there are cues in surface level image changes that can be used to both segment and identify human behavior.

In addition, the problem of determining which subsequences of behavior go together can be seen as an important instance of the more general problem of variable (or feature) discovery and segmentation, a problem that the psychological, machine learning and computer vision literatures have addressed in other domains. Recent work in action segmentation has drawn particularly on both psychological and computational approaches to the problem of segmenting words from continuous speech. There is now a large body of evidence suggesting that both infants and adults can use statistical patterns in spoken language to help segment speech into words (for a partial review, see Gómez & Gerken, 2000). More recently, it was demonstrated that a similar sensitivity to statistical regularities in action sequences could play an important role in action segmentation e.g. (Baldwin, Andersson, Saffran, & Meyer, 2008; Buchsbaum et al., 2009). Similarly, Buchsbaum et al. (2009) were able to successfully adapt a Bayesian model of statistical word segmentation (Goldwater, Griffiths, & Johnson, 2009) to the action domain.

However, like previous computational models of word segmentation, the Buchsbaum et al. (2009) model assumes that the lowest level of segmentation is already known (or pre-labeled). That is, that there is some sort of motion primitive (equivalent to a syllable or phoneme in speech), that can already be recognized as a coherent unit. Since psychological studies demonstrating human action segmentation have suggested that statistical patterns or features in human motion may correlate with segment boundaries at even the lowest level, we would like to see whether action boundaries can be automatically detected directly from video, without pre-existing knowledge of low-level motion units.

Finally, in the computer vision literature, most action recognition work has focused on pre-segmented videos of

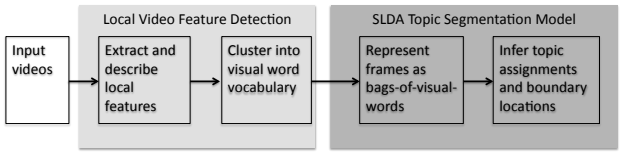

Figure 1: Our video segmentation and recognition pipeline

individual actions, rather than on continuous human action. We would like to better model and understand human action segmentation abilities, by trying to discover whether there are in fact statistical patterns or features in visible human motion that correlate with the action boundaries people identify. Additionally, we would like to develop a computer vision model capable of segmenting (and perhaps recognizing) actions, in videos that contain multiple actions, and more varied actions than in standard computer vision data sets.

Modeling Video Segmentation

We would like to see if statistical patterns in low-level video features correspond to the segmentation structure of human actions, and whether this structure can be discovered directly from video, in an unsupervised fashion. Topic modeling – an approach to classifying language by subject matter – is an appealing starting point for action segmentation because it has already proven successful in modeling simultaneous topic identification and segmentation in transcribed conversations (Purver, Kording, Griffiths, & Tenenbaum, 2006), and because of established parallels between action segmentation and language segmentation (Baldwin et al., 2008; Buchsbaum et al., 2009; Speer, Reynolds, Swallow, & Zacks, 2009).

Recently, Niebles, Wang, and Fei-Fei (2008) demonstrated good action recognition results on several video data sets, by using a combination of local video features and a topic model. In this approach, topic modeling is applied to videos of human action, by constructing a set of "visual words", corresponding to clusters of video features, and re-describing each video as a document composed of these words. An action (or topic) is then a distribution over visual words, and a video (or document) is a distribution over actions (topics).

While the Niebles et al. (2008) results demonstrate the feasibility of applying a topic model to videos of human action, their model was primarily tested on stylized, pre-segmented, individual actions, and the results were not compared to human judgments. Here, we would like to create a model that segments and recognizes video sequences containing multiple actions, performed in more naturalistic settings, and that is consistent with human action segmentation (for an overview of our approach see Figure 1).

Feature Detection and Visual Word Representation

Computer vision approaches using local image and video features have proven surprisingly effective in a variety of scene, object and action recognition tasks (for a recent review see Tuytelaars & Mikolajczyk, 2008). Broadly, these approaches work by searching images or videos for local patches of interest (e.g., blobs, corners, periodic motion), and then creating summary descriptions of the image regions surrounding these interest points, which can then be used as features. They are an appealing choice for this application because they are unsupervised, and work directly with local patches of pixel values, so they do not require person detection or knowledge of pose.

Interest Point Detection In this work, we use the space-time interest point detector introduced by Dollár, Rabaud, Cottrell, and Belongie (2005) and used by Niebles et al. (2008). This feature detector consists of a pair of linearly separable gabor filters which are applied temporally, and convolved with a Gaussian smoothing kernel over the video's spatial dimensions. Space-time interest points correspond to the local maxima of this response function. Dollár et al. (2005) found that this detector responds most strongly to periodic motion, but tends to respond well to any complex motion. The detector has two parameters: a spatial scale s, and a temporal scale t, controlling the size of the pixel volume contributing to each interest point detection. As in Niebles et al. (2008), we run our detector at a single scale, and rely on the visual word vocabulary (described below) to capture any scale variations within a given video.

Feature Descriptor After detecting the video feature locations, we extract a spatial-temporal cube of pixels around each feature center. To obtain a descriptor for each cube, we calculate the brightness gradient and concatenate it to form a vector. This descriptor is then projected to a lower dimensional space using PCA (for a full techinical description of the feature detector and descriptor see Dollár et al., 2005; Niebles et al., 2008). The intuition behind using this type of feature descriptor over the raw pixel intensities, is that it creates relative invariance to slight disparities in appearance (e.g., small changes in motion or lighting).

Visual Words Finally, as in Niebles et al. (2008), we cluster the feature descriptors into v, the vocabulary size, categories, to create the visual word vocabulary. Each detected feature is assigned a cluster label as its "word" type. Cluster assignments are made using the k-means algorithm, and the best (minimum overall distance) set of cluster assignments from 10 runs of the algorithm is used.

Qualitatively, the resulting feature clusters appear intuitively meaningful, often corresponding to motion in a particular direction, or of a specific body part. This results in a set of "utterances" for each processed video, where each utterance is a frame of the input video, each word in the utterance is a video feature occurring at that frame, and each vocabulary item is a cluster of video features with similar descriptors. Example videos of the detector response function, detected features, and vocabulary clusters are available at `http://cocosci.berkeley.edu/videosegdemos.php`

Topic Segmentation Model

Our generative model is the hierarchical Bayesian topic segmentation model presented in Purver et al. (2006). In this model, a meeting is composed of multiple conversations

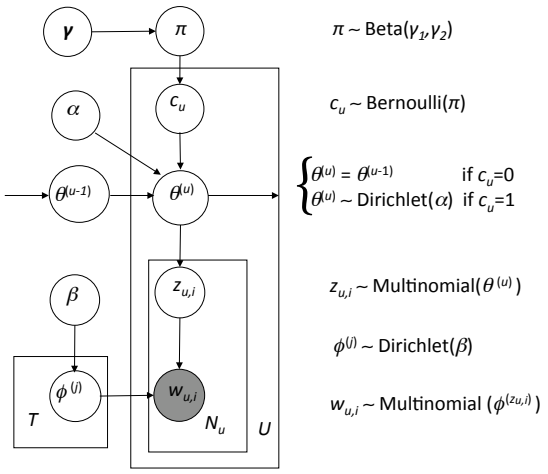

Figure 2: Variable dependencies in the SLDA model

(documents), where each conversation is a series of utterances all generated from the same distribution over topics. As standard, topics are equated with discrete distributions over the set of vocabulary words. The intuition is that a single conversation may contain a number of different topics, but that these topics form a particular coherent group, different from the topics the other conversations range over. Given an unsegmented transcription of a meeting, the model infers the times at which the distribution over topics has shifted, and therefore where the segment boundaries between conversations are.

For our purposes, we can think of a complete video of a series of actions as a meeting, with individual actions being conversations. Similarly, topics are distributions over video words (as in Niebles et al., 2008). Just as certain topics tend to appear together during a conversation and shift between conversations, perhaps certain video topics tend to appear together in actions, and shift between actions.[1]

Figure 2 depicts the entire generative model, which we will call Segmented Latent Dirichlet Allocation (SLDA). The parameter $\pi \sim \text{Beta}(\gamma_1, \gamma_2)$ is the probability that each frame u will begin a new action, in which case $c_u = 1$. Otherwise, $c_u = 0$, and the frame is a continuation of the previous action. All the frames of an action share the same topic distribution, so that $\theta^{(u)} = \theta^{(u-1)}$ if frames u and $u-1$ are part of the same action (that is, if $c_u = 0$), and $\theta^{(u)} \sim \text{Dirichlet}(\alpha)$ if frame u begins a new action (if $c_u = 1$). Within a frame u, the i^{th} visual word, $w_{u,i}$, is generated by the topic $z_{u,i}$, which is drawn from the topic distribution $\theta^{(u)}$ specific to frame u. Each topic j has an associated multinomial distribution $\phi^{(j)}$ over visual words, generated from a Dirichlet(β) distribution. So the visual word $w_{u,i}$ is drawn from the distribution $\phi^{(z_{u,i})}$. The hierarchical topic model has four parameters: α, β, γ_1, and γ_2.

Given an unsegmented video of multiple actions, we can invert the generative model to infer both video topics and segmentation boundaries by sampling from their posterior distribution, conditioned on the observed frames of (unsegmented and unlabeled) video. We can use Gibbs sampling to generate our samples from the posterior distribution, as described in Purver et al. (2006). Each sample is a complete assignment of the c_u and $z_{u,i}$ variables for every frame u in a video, specifying a set of inferred boundaries between video segments of single actions, along with the inferred distribution of topics within each frame of each action.

Implications for Action Segmentation The SLDA topic segmentation model infers boundary locations based on changes in the distribution over topics between frames of video. Therefore, it is sensitive to the patterns of video features that appear together within versus across individual actions, and within actions versus at boundary points. However, the bag-of-visual-words approach means that the model is not aware of spatial or temporal relationships between visual features within an action or a frame. To the extent that individual actions are characterized by particular statistical distributions over combinations of visual features, the model will be successful at discovering action structure. However, if some aspects of action segmentation require knowledge of spatial relations (e.g., between body parts) or of temporal relations (e.g., ordering of sub-goals), and are not also characterized by changes in local features, they will not be captured by this model. The model is therefore a starting point for exploring the amount of representational structure required both to begin parsing human action, and to eventually parse it with adult human accuracy.

Segmentation and Recognition Model Results

We tested our model on three video data sets of everyday human action, described below, and compared the model's segmentation predictions to human judgments (and to ground truth boundary locations, when applicable). For the two data sets with repeating, identifiable actions, we also measured action recognition, by comparing the model's per frame topic assignments to the true action labels for those frames.

All videos were converted to 256 grayscale and 160x120 pixels in preprocessing, to speed feature detection. For each data set, we tested a small range of parameter values for the spatial and temporal scale of feature extraction, and for the size of the visual word vocabulary, $s \in \{1,2,3\}$, $t \in \{2,3,4\}$, $v \in \{25,50,100,250,500,1000\}$, based on values used in previous work (Dollár et al., 2005; Niebles et al., 2008). We also varied the number of topics T used by the SLDA model, to match the number of action types present in the videos. In all cases, SLDA model parameters of $\alpha = 0.1$, $\beta = 0.1$, $\gamma_1 = 1$ and $\gamma_2 = 5$ were used, corresponding to a bias towards having few words per topic, few topics per document, and a prior expectation of a segment boundary approximately every six frames.

For all results, we ran the Gibbs sampler for 24,000 iterations, with a burn-in period of 2000 iterations. Simulated annealing was used during the first 1000 iterations. Per-frame segmentation probabilities were estimated by averag-

[1] This is somewhat different from the approach used by Niebles et al. (2008), where an action is a topic and a complete video is a document. Here, a video contains multiple documents, with an action being a single document, and topics are at the sub-action level.

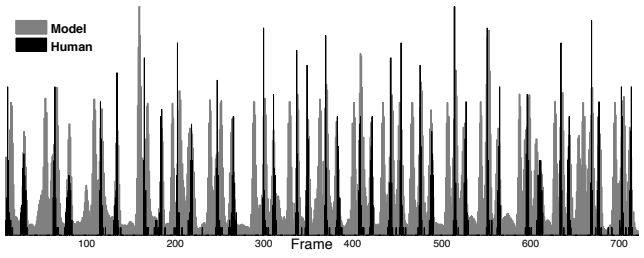

Figure 3: Human boundary judgments and model predictions for the WBD videos

ing 2000 samples, evenly spaced after the burn-in period. Per-frame topic assignments were estimated by first assigning each frame within a single sample to the majority topic over its component words, and then counting the number of times each frame was assigned to each topic, over the 2000 samples.

Segmenting and Identifying Repeated Actions

Recently, Shi, Wang, Cheng, and Smola (2008) presented a supervised, discriminative approach to jointly recognizing and segmenting human actions from video. They created the Walk Bend Draw (WBD) data set, consisting of videos of three subjects, each performing three continuous actions: slow walk, bend body and draw on board. Each subject performed the action sequence 6 times, for a total of 18 videos, averaging 8 seconds each. This data set of short, simple, repeated actions provides a good baseline to test our model. Since human segmentation judgments were not available for this data set, we conducted an experiment to gather ground truth segmentations from human participants.

Participants Participants were 50 English speakers on Amazon's Mechanical Turk.

Stimuli and Procedure For this experiment, we used the WBD data set from Shi et al. (2008), described above. Participants first viewed each movie in its entirety. They were then presented with an interface for stepping through every 5th frame of the video, and were instructed to to step through the movie and mark the frame where the person first begins bending down, and the frame where they first begin drawing on the board (previous work has found boundary judgments gathered by paging through still frames to be reliable, e.g. Meyer et al., 2010; Hard et al., under review). Participants provided boundary judgments for all 18 videos, and video order was randomized across participants. The experiment was run via a custom Flash program.

Experiment Results Data from 7 participants who did not rate all the movies, or who placed the boundaries in incorrect order were discarded. Qualitatively, boundary choices were consistent across participants, and with our own judgments of boundary locations.

Model Settings We ran our model on the concatenation of all 18 WBD videos, using feature extraction and vocabulary parameters $s = 2$, $t = 3$, $v = 250$, and number of SLDA topics, $T = 3$.

Model	Walk	Bend	Draw
SLDA	0.82	0.86	0.87
SVM-SMM	0.78	0.91	0.86

Table 1: Comparison of SLDA action recognition performance, with the supervised model from Shi et al.(2008)

Segmentation Results Model segmentation probabilities were calculated over a 5 frame window, in order to align them with the human data, resulting in human and model boundary judgments at a resolution of 0.17s. Pseudo-Gaussian smoothing with a 0.5s kernel size was applied to the model output, to obtain a more continuous distribution over boundary probabilities. There was a significant correlation between the model's predictions and human boundary judgments, Pearson's correlation coefficient, $r = 0.57$ across all videos, $p < 0.001$. Model predictions and human boundary judgments for the concatenated WBD videos are shown in Figure 3.

Action Recognition Results Although our model does not require that topics have a one-to-one correspondence with actions, we expect few topics per document using our prior, and so the extent to which topic assignments are consistent within actions is a reasonable proxy for action recognition.

To measure per-frame recognition accuracy, we first assigned each video frame to the most common topic across samples. We then computed the most common topic for each action type, and the proportion of frames of each action type assigned to this topic (we used hard boundaries located at the local maxima of human segmentation judgments for these calculations). Results are shown in Table 1. Recognition accuracy was high, and comparable to that achieved by Shi et al. (2008) using a supervised approach.

Hierarchical Statistically Grouped Actions

The second test of our model used the video corpus from Experiment 1 of Buchsbaum et al. (2009). This video corpus was designed to replicate the structure of artificial language learning experiments. Since previous work (Baldwin et al., 2008; Buchsbaum et al., 2009) has established that adults are able to recognize artificial *actions* – triplets of smaller motion elements, grouped only by their statistical regularities – this corpus is an interesting test case for whether our model will also pick up on this hierarchical grouping of motions.

In this corpus, 12 individual video clips of object-directed motions (referred to as *small motion elements* or SMEs in previous work) were used to create four *actions* composed of three SMEs each. A 25 minute corpus was created by randomly choosing actions to add to the sequence, resulting in 90 appearances of each action and 30 appearances of each transition between pairs of actions. After viewing the corpus, participants were asked to rate individual *actions*, as well as *part-action* and *non-action* comparison stimuli, on how meaningful and coherent they felt each combination of three SMEs was. As is standard in this genre of experiments, a *part-action* was a combination of three SMEs that appears across a transition (e.g., the last two SMEs from the first action and the first SME from the second action), and a

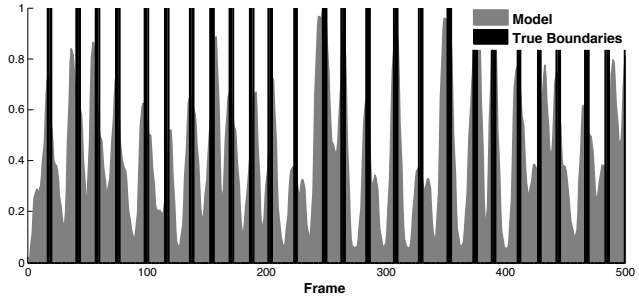

Figure 4: True boundary locations and model predictions for the first 500 frames of the Buchsbaum et al. (2009) corpus

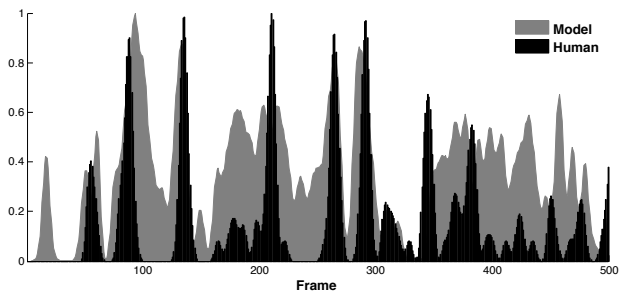

Figure 5: Human "fine" boundary data and model predictions. First 500 frames of Zacks, Braver, et al.(2001) video

non-action was a combination of three SMEs that never appear together in the exposure corpus. Buchsbaum et al. (2009) found that participants rated *actions* as more coherent than *part-actions* and *non-actions*, suggesting they perceive them as distinct, meaningful action segments.

Model Settings We down-sampled the exposure corpus from 30 to 5 fps, and ran our model on the complete 25 minute video, using feature extraction and vocabulary parameters $s = 2$, $t = 2$, $v = 500$, and number of topics, $T = 4$.

Segmentation Results Since the corpus was artificially assembled, the true locations of transitions between actions are known. For this analysis, we consider the 2 frames on either side of a boundary as the ground-truth frames. Pseudo-Gaussian smoothing with a 0.8s kernel size was applied to the model output, to obtain a more continuous distribution over boundary probabilities.

There was a significant correlation between the model's boundary probabilities and human boundary judgments, Pearson's correlation coefficient, $r = 0.49$, $p < 0.001$. The first 500 frames of model predictions and ground-truth boundaries are shown in Figure 4.

Action Recognition Results Per-frame recognition accuracy for actions was computed as for the WBD data set. Recognition accuracy for non-actions and part-actions was computed in a similar manner, by looking at the most common topic assignments for the component motions (without requiring these motions to appear sequentially). Results are shown in Table 2. Overall, the model's per-frame topic assignments were very similar to human coherence judgements from Experiment 1 of Buchsbaum et al. (2009).

Segmenting Naturalistic Action

The previous two data sets consist of a set of repeated actions, performed in a relatively uncluttered environment. To look at whether our model can use low-level features to find structure in more realistic action, at multiple hierarchical levels, we tested it on a video of everyday activity, from Zacks, Braver, et al. (2001). The video is 336 seconds long, and shows a person making a large bed. Sixteen participants provided online segmentation data for this video, at two levels of action granularity ("coarse" and "fine" boundaries respectively).

Table 2: Comparison of SLDA action recognition, with human coherence ratings from Buchsbaum et al. (2009)

	Actions	Part-Actions	Non-Actions
SLDA	0.67	0.54	0.43
People	0.63	0.53	0.46

Model Settings We down-sampled the exposure corpus from 30 to 5 fps, and ran our model on the complete 336s video. Since human boundaries were given in milliseconds, each human boundary was re-assigned to the closest following frame, resulting in human and model boundary judgments at a resolution of 0.20s. Previous work on this type of naturalistic video binned segmentation judgments at a 1s resolution (e.g., Zacks, Braver, et al., 2001; Zacks et al., 2009), so Pseudo-Gaussian smoothing with a 1s kernel size was applied to both human boundaries and model predictions. Since participants provided online boundary judgments, there might be a response time delay between the true boundaries, and human judgments. To account for this possibility, we looked at correlations between the model and human segmentation, with the model shifted between 0 and 5 frames (0 to 1s).

Fine Segmentation Results There was a significant correlation between the model's boundary probabilities and human fine grain boundary judgments, Pearson's correlation coefficient, $r = 0.31$, $p < 0.001$, using feature extraction and vocabulary parameters $s = 2$, $t = 3$ $v = 250$, shift $= 0$, and number of topics, $T = 10$. The first 500 frames of model predictions and human "fine" boundaries are shown in Figure 5.

Coarse Segmentation Results There was a significant correlation between the model's boundary probabilities and human coarse grain boundary judgments, Pearson's correlation coefficient, $r = 0.35$, $p < 0.001$, using feature extraction and vocabulary parameters $s = 3$, $t = 2$ $v = 250$, shift $= 3$, and number of topics, $T = 3$. The first 500 frames of model predictions and human "coarse" boundaries are shown in Figure 6. The greater lag between the model and participant responses seen here is consistent with previous work suggesting that coarse grain boundaries may take longer to process than fine grain boundaries (e.g Meyer et al., 2010; Hard et al., under review). The smaller number of topics is consistent with there being fewer, larger "coarse" versus "fine" actions.

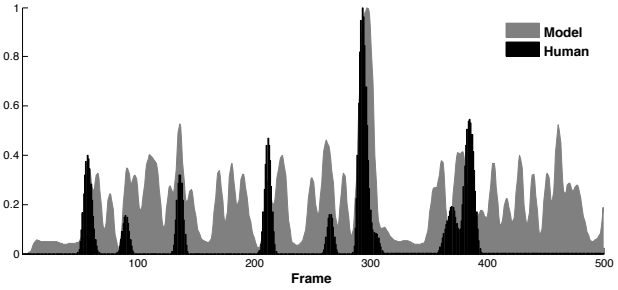

Figure 6: Human "coarse" boundary data and model predictions. First 500 frames of Zacks,Braver, et al.(2001) video

Discussion

Across three sets of videos, our model's action segmentation predictions were significantly correlated with human boundary judgments and ground-truth boundary locations. The model was able to discover relevant local visual features, and statistical patterns over those features, for all videos, even though they contained a variety of actors, were filmed from different angles, and encompassed actions performed at multiple time-scales and hierarchical levels. Despite having no a priori knowledge of person location or human pose, the correlations between our model's boundary predictions and human judgments were comparable to those found in previous work looking at pre-selected movement features of tracked people and body parts (Zacks et al., 2009; Meyer et al., 2010). In the case of the Buchsbaum et al. (2009) data set, our model was able to make reasonable boundary predictions, even though the "actions" were artificially assembled, and were defined only by co-occurence probabilities of the lower level motion units. Similarly, our model made meaningful boundary predictions for the Zacks, Braver, et al. (2001) "bed" video, for multiple hierarchical levels of segmentation.

In addition, our model was able to successfully identify repeated actions within the videos, assigning consistent topic labels to occurrences of the same action. On the WBD data set, the performance of our unsupervised approach was comparable to that of earlier work using labeled training examples (Shi et al., 2008), and on the Buchsbaum et al. (2009) corpus, our model's performance was qualitatively very similar to human judgments of action coherence.

This model does not make explicit distinctions between people and objects, or foreground and background. It also does not look for pre-defined motion features, such as speed or acceleration. Therefore, to the extent that our approach was successful, it indicates that a portion of human action structure is discoverable simply by attending to locally salient visual cues within a broader visual scene. This suggests that attending to such cues is a potential mechanism by which learners (whether infants or computer vision algorithms) might bootstrap their way into action parsing.

Finally, our model did not fully capture human boundary judgments, and it performed more poorly on the less repetitive, more naturalistic action sequence. Differences between our model and human performance indicate areas where additional knowledge may be required to fully parse observed actions. We would like to explore the extent to which adding representational (e.g., human pose), contextual (e.g., object knowledge) and social (e.g., intentions and goals) information improve results in future work.

Acknowledgments. We thank Jeff Zacks, Dare Baldwin, Meredith Meyer, Misha Shashkov, and Andy Horng. This work was supported by an NSF Graduate Research Fellowship and grant FA-9550-10-1-0232 from the Air Force Office of Scientific Research.

References

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*, 1382-1407.

Baldwin, D., & Baird, J. (2001). Discerning intentions in dynmic human action. *Trends In Cognitive Sciences*, *5*, 171-178.

Baldwin, D., Baird, J., Saylor, M., & Clark, A. (2001). Infants parse dynamic human action. *Child Development*, *72*, 708-717.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. *Proc. of the 31st Annual Conference of the Cognitive Science Society*.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. *VS-PETS*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178-186.

Hard, B., Recchia, G., & Tversky, B. (under review). The shape of action.

Hard, B., Tversky, B., & Lang, D. (2006). Making sense of abstract events: Building event schemas. *Memory and Cognition*, *34*, 1221-1235.

Meyer, M., DeCamp, P., Hard, B. M., & Baldwin, D. A. (2010). Assessing behavioral and computational approaches to naturalistic action segmentation. *Proc. of the 33nd Annual Conference of the Cognitive Science Society*.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847-862.

Niebles, J., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial temporal words. *International Journal of Computer Vision*, *79*(3), 299-318.

Purver, M., Kording, K. P., Griffiths, T. L., & Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of COLING/ACL*.

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition andDevelopment*, *8*(1), 113-128.

Shi, Q., Wang, L., Cheng, L., & Smola, A. (2008). Discriminative human action segmentation and recognition using a semi-markov model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Speer, N., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science*(20), 989-999.

Tuytelaars, T., & Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, *3*(3).

Zacks, J. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*, 979-1008.

Zacks, J., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651-655.

Zacks, J., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*, 201-216.

Zacks, J., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: segmentation of narrative cinema. *Frontiers in Human Neuroscience*, *4*(1-15).

Zacks, J., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29–58.