Behavior Recognition in Video with Extended Models of Feature Velocity Dynamics

Ross Messing[1]
[1]Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
{rmessing, cpal}@cs.rochester.edu

Christopher Pal[1,2]
[2]Département de génie informatique et génie logiciel
École Polytechnique de Montréal
Montréal, QC, H3C 3A7, Canada

Abstract

Investigations of human perception have shown that non-local spatio-temporal information is critical and often sufficient for activity recognition. However, many recent activity recognition systems have been largely based on local space-time features and statistical techniques inspired by object recognition research. We develop a new set of statistical models for feature velocity dynamics capable of representing the long term motion of features. We show that these models can be used to effectively disambiguate behaviors in video, particularly when extended to include information not captured by motion, like position and appearance. We demonstrate performance surpassing and in some cases doubling the accuracy of a state-of-the-art approach based on local features. We expect that long range temporal information will become more important as technology makes longer, higher resolution videos commonplace.

Introduction

Activity recognition is an important problem in computer vision. Like many vision problems, it is difficult to explicitly characterize what information is most useful for a range of domains. Despite this difficulty, activity recognition has many applications. In addition to applications in security and surveillance, health monitoring systems also require activity recognition. An important example of these health monitoring systems are assisted cognition systems designed to monitor patients unobtrusively, in order to ensure their mental and physical health either in their home or an extended care facility. Additionally, these systems can provide critical, otherwise unavailable information to family members or physicians making a diagnosis. With an aging population, and without a concurrent increase in healthcare workers, techniques that could both lessen the burden on caregivers and increase quality of life for patients by unobtrusive monitoring are very important.

As in object recognition, bag of visual words approaches currently dominate the statistical activity recognition literature.While almost every approach uses motion in one way or another, most make strong restrictions on the spatio-temporal range of motion that they consider. In particular, few approaches consider velocity outside of a space-time window around an interest point. Human performance suggests that considering more global temporal information could yield improved performance.

Background

Researchers have taken a variety of approaches to activity or human behavior recognition in video. Approaches generally fall into one of two categories - approaches based on local appearance (features), and approaches based on global appearance.

Local appearance

In object recognition, techniques using spatially local features can often produce state-of-the-art results. Many approaches to object recognition involve extracting features such as Lowe's SIFT descriptors (Lowe 2004) at the space-time peaks of an interest point operator or through various forms of dense location sampling of the scene. Frequently, features are then discretized into a codebook, and a discriminative or generative model is fit to the data. It is common for simple generative bag of visual words models to be used, ignoring the spatial relationships between features as this can yield impressive results and fast implementations.

This popular approach in object recognition has been generalized to action recognition in video. Dollár et al. (Dollár et al. 2005) use cuboids in space-time extracted around space-time interest points to recognize activities. These cuboids are processed to yield features (normalized illumination, optical flow, or brightness gradient at each space-time location), the features are discretized into a codebook, and an activity is modeled as a distribution over codebook entries. Recently, a number of vision researchers have embraced more sophisticated Bayesian techniques (like latent Dirichlet allocation (Blei, Ng, and Jordan 2003)). For example, Niebles et al. (Niebles, Wang, and Fei-Fei 2006) used Dollár et al.'s spatio-temporal cuboids as basic features in an unsupervised hierarchical bag-of-features model.

Global appearance

Some activity recognition work has relied primarily on global rather than local information. For example, Ramanan and Forsyth (Ramanan and Forsyth 2003) infer an appearance-based generative model of pose and learn a 2D to 3D mapping with motion capture training data. A novel item is compared to clustered training data in 3D-pose space. This is typical of the highly structured model-based approach to human tracking. While this approach can infer a great deal of useful information, its primary purpose is high-level pose modeling. Pose can be helpful in activity recognition, but motion information can potentially be just as useful for activity recognition, and these models do not generally use motion beyond inferring high-level pose.

A few recent activity recognition methods directly generalize local appearance methods to global appearance. For example, Niebles and Fei-Fei (Niebles and Fei-Fei 2007) extend Niebles et al.(Niebles, Wang, and Fei-Fei 2006) by adding a constellation model characterizing the relative spatial layout of ”parts”. In their model, each part is a Bag-of-Words, as in their original work.

Global Motion

While much of the work reviewed so far has revolved around a combination of local appearance and motion, human activity recognition can be facilitated by global motion perception in the absence of useful appearance information. Johansson (Johansson 1973) showed that point-light displays of even very complicated, structured motion were perceived easily by human subjects . Most famously, Johansson showed that human subjects could correctly perceive “point-light walkers”, a motion stimulus generated by a person walking in the dark, with points of light attached to the walker’s body. This leads us to speculate that motion aloneis sufficient to recognize activities in the absence of local appearance information. Naturally, most work on video sequences analyzes change over time. One way to do this is optical flow, which provides a dense measure of change for every point in the image between every adjacent pair of frames. Despite the density of optical flow, it can be a shallow feature, conveying a great deal of noise along with a limited amount of signal. Some way of characterizing motion at particularly interesting points, analogous to the point-lights on Johansson’s walkers, could convey almost as much signal, with much less noise.

The interest point centered, space-time cubiod features of Dollár et al.(Dollár et al. 2005) represent a move towards the point lights of Johnsson; however, there are a number of differences. The space-time cuboid approach only deals with local motion that falls fully within the space-time window captured by the feature descriptor and it thus has no way of dealing with much longer range motion. Savarese et al. (Savarese et al. 2008) have extended this approach by adding spatio-temporal correlatons to the bag of features, which capture the relative positions of features. This approach works well with a bag-of-features model, but fails to capture the information contained in the extended motion of any particular feature. Madabhushi and Aggarwal (Madabhushi and Aggarwal 1999) described a system using the motion trajectory of a single feature, the center of the head, to distinguish between activities. While the system had limited success, the ability to use extremely sparse feature motion for activity recognition shows how informative feature motion can be. In the spirit of this an other trajectory modeling based approaches, we seek to determine whether increasing the quality and quantity of non-local motion information could provide a robust source of information for activity recognition. As such, we develop a model capable of using the full velocity history of a tracked feature.

Much of the previous trajectory tracking work has often been based on object centroid or bounding box trajectories. In a more sophisticated extension of such approaches, Rosales and Sclaroff (Rosales and Sclaroff 1999) used an extended Kalman filter to help keep track of bounding boxes of segmented moving objects and then use motion history images and motion energy images as a way of summarizing the trajectories of tracked objects. Madabhushi and Aggarwal (Madabhushi and Aggarwal 1999) the trajectories of heads tracked in video and built models based on the mean and covariance of velocities. Our approach to analyzing feature trajectories differs dramatically from these previous methods in that we use dense clouds of KLT (Lucas and Kanade 1981) feature tracks. In this way our underlying feature representation is also much closer to Johansson’s point lights.

Position

While feature motion is often informative, and clearly plays a primary role for humans perceiving activities, it is not perceived in a vacuum even in point-light displays, where the relative positions of the point-lights are also available to the viewer. Many extensions to bag-of-features models in vision focus on weakening the geometric assumptions made by that model. In addition to Niebles and Fei-Fei’s hierarchical constellation model(Niebles and Fei-Fei 2007), and Wong et al.’s extended topic model (Wong, Kim, and Cipolla 2007), other vision work outside of activity recognition has used this approach to add spatial structure to a model that originally had none. The best known example of this might be Fergus et al.’s constellation model (Fergus, Perona, and Zisserman 2003) for object recognition. Clearly, augmenting a feature with position provides a powerful, tractable way to incorporate position into a bag-of-features approach.

Appearance

In addition to position, appearance information can be extremely useful in disambiguating activities. While human subjects can recognize actions performed by Johansson’s point-light walkers (Johansson 1973), it is even easier to identify activities when shown unrestricted videos. Local appearance patches have become the dominant approach to object recognition, and direct extensions of these approaches based on local space-time patches make up a large part of the statistical activity recognition literature.

Activities to Recognize

Activity recognition has no accepted standard dataset. It is difficult to compare the performance of systems meant to solve the same, or similar tasks. This is particularly troublesome because many researchers generate their own datasets, and it is not always clear what assumptions went into their generation. The most popular activity recognition dataset is the KTH dataset (Schuldt, Laptev, and Caputo

2004), used by a number of systems (Dollár et al. 2005; Wong, Kim, and Cipolla 2007; Savarese et al. 2008; Niebles, Wang, and Fei-Fei 2006). This dataset consists of low-resolution (160 × 120 pixels) video clips taken by a non-moving camera of 25 subjects performing each of 6 actions (walking, clapping, boxing, waving, jogging and running) in several background-variation conditions. While this dataset has an appealingly large amount of data for each activity, it is not really suitable to test an assisted cognition system. First, there is a lot of variation in the background between video sequences, while we can assume an almost constant background in an assisted cognition context. Second, the background in the KTH dataset is uninformative, while background can provide a great deal of contextual information for activity an assisted cognition task. In addition, the dataset consists largely of full-body activities that involve no objects, while object interaction can provide a valuable source of information for the kind of activities of daily living that an assisted cognition system must deal with. Lastly, the low resolution of this dataset may compromise the quality of the feature trajectories extracted, both in velocity resolution, and track duration.

In order to evaluate an activity recognition system for an assisted cognition task, we construct a new dataset to overcome the shortcomings we have identified in the KTH dataset.

A Model of Feature Velocity Dynamics and Extensions

We present a system that extracts a set of feature trajectories from a video sequence, augments them with information unrelated to motion, and uses a model of how those trajectories were generated to classify the trajectory sets.

Method

Flow extraction Given a video sequence, we extract feature trajectories using Birchfield's implementation (Birchfield 1996) of the KLT tracker (Lucas and Kanade 1981). This system finds interest points where both eigenvalues of the matrix of the image gradients are greater than a threshold, and tracks them by calculating frame-to-frame translation with an affine consistency check. In order to maximize the duration of our feature trajectories, and thus capitalize on their descriptive power, we did not use the affine consistency check (which ensures that the features have not deformed). This increased the amount of noise in the feature trajectories, but also the ability to capture non-rigid motion. As described by Baker and Matthews (Baker and Matthews 2004), the tracker finds the match that minimizes

$$\sum_{x}[W(x;p) - T(x)]^2 \qquad (1)$$

where T is the appearance of the feature to be matched in the last frame, x is the position in the template window, W are the set of transformations considered (in this case, translation) between the last frame and the current one, and p are the parameters for the transformation. We tracked 500 features at a time, replacing lost features with the best new features the tracker could find. We call each feature's quantized velocity over time its "Feature Trajectories", and use this as our basic feature. We emphasize that we are unaware of other work on activity recognition using a motion description with as long a temporal range as the technique presented here. Uniform quantization is done in log-polar coordinates, with 8 bins for direction, and 5 for magnitude. The video was captured at 30 frames per second, and its initial temporal resolution was maintained. We limited the velocity magnitude to 10 pixels per frame, which in our high-resolution dataset corresponds almost exclusively to noise. This limit meant that for an individual feature's motion from one frame to the next, any motion above the threshold was not ignored, but that feature's motion on that frame was placed in the bin for the greatest motion magnitude in the appropriate direction. We also only considered feature trajectories lasting at least a third of a second. This heuristic reduces noise while eliminating little of the extended velocity information we seek to capture. Examples of the trajectories being extracted can be seen in Figure 1.

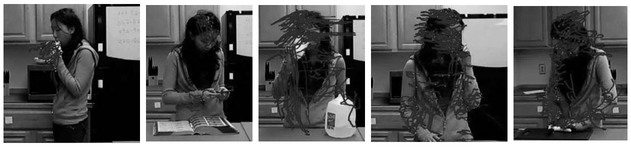


Figure 1: Example trajectories from the following activities (from left to right): eating a banana, dialing a phone, drinking water, answering the phone and chopping a banana

Figure 2: A background image from the activities of daily living data set

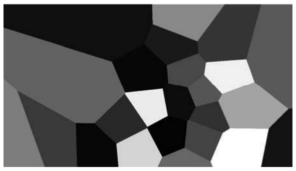


Figure 3: Codebook-generated pixel map of birth-death locations

Feature Augmentations Features were augmented by location information indicating their initial and final positions (their "birth" and "death" positions). To obtain a given feature trajectory's birth and death, we use k-means clustering ($k = 24$)to form a codebook for both birth and death (independently), and assign a feature's birth and death to be it's nearest neighbor in the codebook. The results of this codebook can be seen in Figure 3, which was generated from activities in the scene shown in Figure 2.

In addition to this absolute position information, features are also augmented by relative position information. We use birth and death positions of the feature relative to the position of an unambiguously detected face, if present. Like birth and death, these positions are clustered by k-means ($k = 100$) into a codebook, and a feature position relative to the face is assigned to it's nearest neighbor in the codebook.

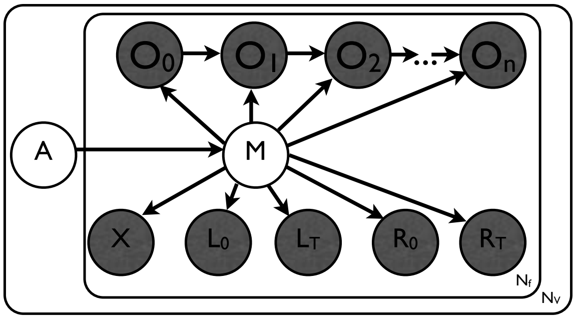


Figure 4: Graphical model of a set of actions (Dark circles denote observed variables).

Lastly, we augment features with local appearance information. We extract a 21 × 21 pixel patch around the initial position of the feature, normalized to the dominant orientation of the patch, calculate the horizontal and vertical gradients of the oriented patch, and use PCA-SIFT (Ke and Sukthankar 2004) to obtain a 40-dimensional descriptor. These descriptors are clustered by k-means into a codebook ($k = 400$), and each feature's patch is assigned to it's nearest neighbor in the codebook.

Activity recognition We model activities as a weighted mixture of bags of augmented trajectory sequences. Each activity model has a distribution over 100 shared mixture components. By analogy to supervised document-topic models, each activity is a document class with a distribution over mixture components (analogous to topics). Continuing the analogy, each mixture component has a distribution over features. To make this more explicit, each mixture component has a discrete distribution over a feature's augmentations, and over it's trajectory. Each trajectory is generated by a system making a Markov assumption over velocity. This is analogous to the human subject's internal model of a person walking while viewing the Johanssson point-light walker stimulus.

The mixture of Markov models is implemented by evaluating the observed velocity sequence with each mixture model's velocity transition matrix. Birth and death are treated as additional observations which are generated by the mixture component.

Inference Our model for actions is shown in Figure 4. Each action is a mixture model. Each instance of an action (video clip) is treated as a bag of augmented trajectories, where a feature's observations are its discretized velocity at each time step, and its augmentations. A is the action variable, which can take a value for each distinct action the system recognize. M is the mixture variable, indicating one of the action's mixture components. L_0 and L_T are the birth and death location variables respectively, R_0 and R_T are the positions relative to the face at birth and death, and X is the feature's appearance. These augmentation variables depend only on the mixture component. O_n denotes the observation at time n. N_f denotes the number of augmented features in the video sequence, and N_v denotes the number of video sequences.

Our joint probability model for an action is:

$$\begin{aligned} &P(A, M, L_0, L_T, R_0, R_T, X, S, O) = \\ &P(A) \prod_f^{N_f} \prod_i^{N_m} P(M_f^i|A) P(L_{0,f}^i|M^i) P(L_{T,f}^i|M^i) \\ &P(R_{0,f}^i|M^i) P(R_{T,f}^i|M^i) P(X_f^i|M^i) P(O_{0,f}^i|M^i) \\ &\prod_{t=1}^{T} P(O_{t,f}^i|O_{t-1,f}^i) \qquad (2) \end{aligned}$$

where N_m is the number of mixture components, N_f is the number of features, and T is the number of observations in a given feature trajectory. For a feature tracked over 20 frames, t would have 19 disctinct values, since each observation is given by the difference between the feature's location between two frames. Note that all random variables are discrete, and all component distributions in the joint distribution are multinomial.

We train the model using Expectation Maximization (EM) (Bilmes 1997), and classify a novel video sequence D by finding the action A that maximizes $P(A|D) \propto P(D|A)P(A)$. We assume a uniform prior on action, so $\text{argmax}_A P(A|D) \propto \text{argmax}_A P(D|A)$.

We also give Dirichlet priors to each of the model parameters in order to smooth over the sparse data. These priors are chosen to be equivalent to having seen each value of every augmentation once from each mixture component, and each next velocity state (within the same mixture component) once for each velocity state in each mixture component.

Evaluation

In order to assess the degree that augmented motion trajectories characterize activities, we generated a data set involving activities of daily living.

Activity Categories

Several different activities were tested. These activities were chosen to be common activities of daily living, each involving different kinds of motion.

The full list of activities is: answering a phone, dialing a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware.

These activities were each performed three times by five different people. By using people of different shapes, sizes, genders, and ethnicities, we hoped to ensure sufficient appearance variation to make the individual appearance features generated by each person different. We also hoped to ensure that our activity models were robust to individual variation.

The scene was shot from about two meters away by a tripod-mounted Canon Powershot TX1 HD digital camera. Video was taken at 1280 × 720 pixel resolution, at 30 frames per second. An example background image is shown in Figure 2.

The total number of features extracted per activity sequence varied between about 700 and 2500, with an average of over 1400 features per sequence. The mean duration of the trajectories was over 150 frames. Video sequences lasted between 10 and 60 seconds, terminating when the activity was completed.

While our dataset was captured from a static camera, only our absolute birth and death position augmentations lack a significant degree of view invariance. The other augmentations, and the feature trajectory model itself, maintain the same view invariance as other feature-based approaches to activity recognition.

Our evaluation consisted of training on all repetitions of activities by four of the five subjects, and testing on all repetitions of the fifth subject's activities. This leave-one-out testing was averaged over the performance with each left-out subject.

Comparison

To evaluate the performance of our algorithm on our novel dataset, we implemented Dollár et al.'s spatio-temporal cuboid-based discriminative classifier (Dollár et al. 2005). This system finds the local maxima of a spatio-temporal interest point detector, extracts local space-time features ("cuboids") around those peaks, describes them and reduces their dimensionality using a technique similar to PCA-SIFT (Ke and Sukthankar 2004). These lower-dimensional features are then clustered into a codebook, each training and testing movie is described as a histogram of these discretized features, and the histograms are used as training and testing inputs for a discriminative classifier (an SVM).

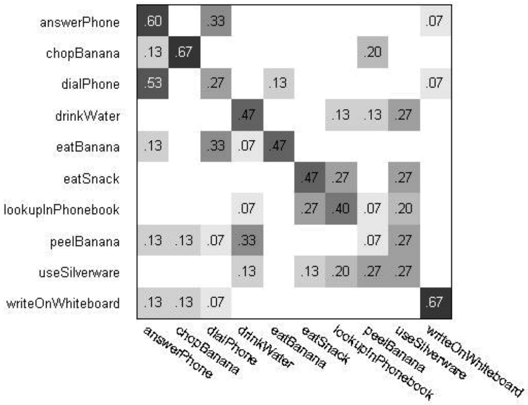
We attempted to use parameter settings similar to those used by Dollár et al. (Dollár et al. 2005) and Niebles et al. (Niebles, Wang, and Fei-Fei 2006). We set cuboid spatial and temporal scale (σ and τ) to 2 and 3, respectively. This meant that we extracted cuboids of size $13 \times 13 \times 19$ pixels around the interest point detector peaks. We use a gradient descriptor, and reduced the dimensionality of the descriptors to 100. We formed a codebook of 500 elements from a sample of 60,000 features from the training set, although we found almost identical performance across a range of codebook sizes.

We trained both systems on all repetitions of activities by four of the five subjects, and tested on all repetitions of the fifth subject's activities. This leave-one-out testing was averaged over the performance on each left-out subject.

Results

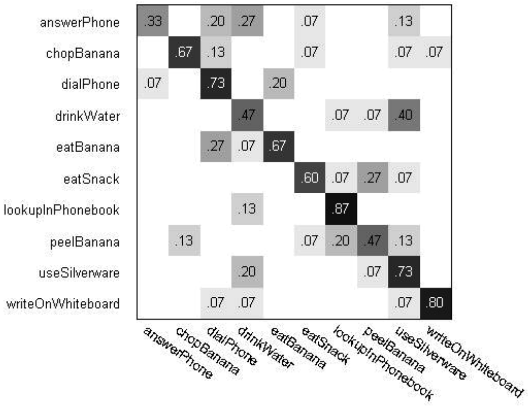
The confusion matrix for Dollár et al.'s system is shown in Figure 5, while the confusion matrix for our augmented feature trajectory model is shown in Figure 8.

In order to investigate the relative contributions of feature trajectory and augmentation information, we also trained a mixture model using only feature trajectories, without augmentations. Everything else about the calculations remained the same as in the augmented feature trajectory model. The performance of the unaugmented model is shown in Figure



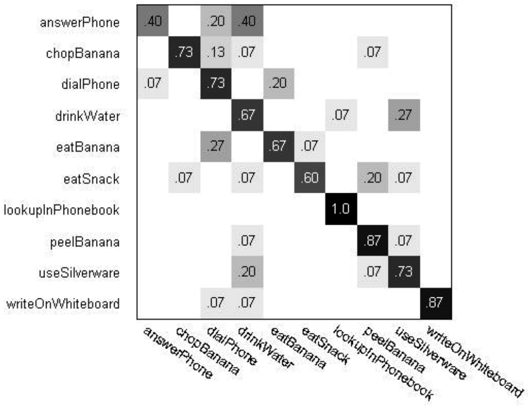
	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookupInPhonebook	peelBanana	useSilverware	writeOnWhiteboard
answerPhone	.60		.33							.07
chopBanana	.13	.67						.20		
dialPhone	.53		.27		.13					.07
drinkWater				.47			.13	.13	.27	
eatBanana	.13		.33	.07	.47					
eatSnack						.47	.27		.27	
lookupInPhonebook				.07		.27	.40	.07	.20	
peelBanana	.13	.13	.07	.33				.07	.27	
useSilverware				.13		.13	.20	.27	.27	
writeOnWhiteboard	.13	.13	.07							.67

Figure 5: Confusion matrix for the histogram of spatio-temporal cuboids behavior recognition system. Overall accuracy is 43%. Zeros are omitted for clarity



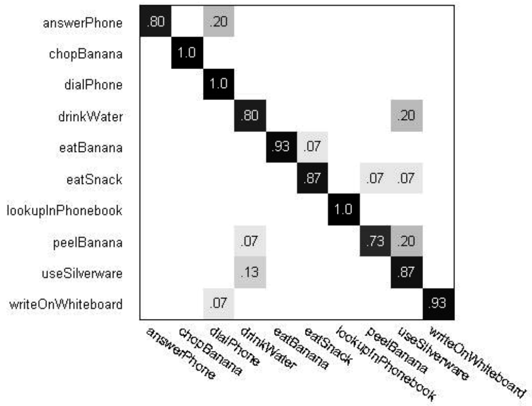
	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookupInPhonebook	peelBanana	useSilverware	writeOnWhiteboard
answerPhone	.33		.20	.27		.07			.13	
chopBanana		.67	.13			.07			.07	.07
dialPhone	.07		.73		.20					
drinkWater				.47			.07	.07	.40	
eatBanana			.27	.07	.67					
eatSnack						.60	.07	.27	.07	
lookupInPhonebook				.13			.87			
peelBanana		.13				.07	.20	.47	.13	
useSilverware				.20				.07	.73	
writeOnWhiteboard			.07	.07					.07	.80

Figure 6: Confusion matrix using only feature trajectory information without augmentations. Overall accuracy is 63%. Zeros are omitted for clarity



	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookupInPhonebook	peelBanana	useSilverware	writeOnWhiteboard
answerPhone	.40		.20	.40						
chopBanana		.73	.13	.07				.07		
dialPhone	.07		.73		.20					
drinkWater				.67			.07		.27	
eatBanana			.27		.67	.07				
eatSnack		.07		.07		.60		.20	.07	
lookupInPhonebook							1.0			
peelBanana				.07				.87	.07	
useSilverware				.20				.07	.73	
writeOnWhiteboard			.07	.07						.87

Figure 7: Confusion matrix for the augmented feature trajectory activity recognition system, without the absolute birth and death position augmentations. Overall accuracy is 72%. Zeros are omitted for clarity



	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookupInPhonebook	peelBanana	useSilverware	writeOnWhiteboard
answerPhone	.80		.20							
chopBanana		1.0								
dialPhone			1.0							
drinkWater				.80					.20	
eatBanana					.93	.07				
eatSnack						.87		.07	.07	
lookupInPhonebook							1.0			
peelBanana				.07				.73	.20	
useSilverware				.13					.87	
writeOnWhiteboard			.07							.93

Figure 8: Confusion matrix for the augmented feature trajectory activity recognition system. Overall accuracy is 89%. Zeros are omitted for clarity

6. Because the information in the absolute birth and death augmentation may be more a result of our static background and camera than the semantic region labels we hope it captures, we trained the model using all augmentations but the birth and death. Figure 7 shows the performance of this test.

As the results show, even unaugmented feature trajectories significantly outperform spatio-temporal cuboids on our dataset of complicated, high-resolution activities, and augmented feature trajectories significantly outperform them.

Discussion

Activity recognition is a difficult problem with a rich, noisy dataset. Many current statistical approaches to activity recognition in video directly generalize techniques from object recognition across time. Their prominence in the litterature speaks to both the intuitiveness of this extension, where a video is just a stack or cube of images, and the power of these statistical models, which generalize from two-dimensional images to three-dimensional video without requiring temporal structure beyond local feature extraction. That said, these feature-techniques often fail to exploit non-local temporal information, and a stack of images, while intuitive, may not be the most useful way to think about video. Videos of activities have rich global temporal structure that is potentially even more informative than the spatial structure in images. This information is often thrown away by activity recognition systems inspired by the success of the spatial-independence assumptions of bag-of-features object recognition systems. We have shown that a relatively simple system using that global motion information can outperform a system using more sophisticated local features and powerful discriminative methods. As high resolution recording devices become more and more common, video will become both richer and noisier. This will lead to an increase in both the quality of the global motion information, and the quantity of uninformative local features. By acknowledging that the structure of video is more than just a stack of images, we can develop models and features that better exploit it's global sequential structure.

References

Baker, S., and Matthews, I. 2004. Lucas-kanade 20 years on: A unifying framework. *IJCV* 56(3):221 – 255.

Bilmes, J. 1997. A gentle tutorial on the EM algorithm. Technical Report ICSI-TR-97-021, University of Berkeley.

Birchfield, S. 1996. Klt: An implementation of the kanade-lucas-tomasi feature tracker.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

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

Fergus, R.; Perona, P.; and Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 264–271.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics* 14:201–211.

Ke, Y., and Sukthankar, R. 2004. Pca-sift: a more distinctive representation for local image descriptors. In *CVPR04*, II: 506–513.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Lucas, B. D., and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 674–679.

Madabhushi, A., and Aggarwal, J. K. 1999. A bayesian approach to human activity recognition. In *VS '99: Proceedings of the Second IEEE Workshop on Visual Surveillance*, 25. Washington, DC, USA: IEEE Computer Society.

Niebles, J. C., and Fei-Fei, L. 2007. A hierarchical model model of shape and appearance for human action classification. In *IEEE CVPR*.

Niebles, J. C.; Wang, H.; and Fei-Fei, L. 2006. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*.

Ramanan, D., and Forsyth, D. A. 2003. Automatic annotation of everyday movements. In *Neural Information Processing Systems (NIPS)*.

Rosales, R., and Sclaroff, S. 1999. Trajectory guided tracking and recognition of actions. Technical report, Boston University, Boston, MA, USA.

Savarese, S.; Pozo, A. D.; Niebles, J. C.; and Fei-Fei, L. 2008. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*.

Schuldt, C.; Laptev, I.; and Caputo, B. 2004. Recognizing human actions: A local svm approach. In *ICPR*, 32–36.

Wong, S.-F.; Kim, T.-K.; and Cipolla, R. 2007. Learning motion categories using both semantic and structural information. In *CVPR*.