The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities

Hilde Kuehne*
Fraunhofer FKIE
Bonn, Germany
hildegard.kuehne@fkie.fraunhofer.de

Ali Arslan
Brown University
Providence, RI
ali_arslan@brown.edu

Thomas Serre
Brown University
Providence, RI
thomas_serre@brown.edu

Abstract

This paper describes a framework for modeling human activities as temporally structured processes. Our approach is motivated by the inherently hierarchical nature of human activities and the close correspondence between human actions and speech: We model action units using Hidden Markov Models, much like words in speech. These action units then form the building blocks to model complex human activities as sentences using an action grammar.

To evaluate our approach, we collected a large dataset of daily cooking activities: The dataset includes a total of 52 participants, each performing a total of 10 cooking activities in multiple real-life kitchens, resulting in over 77 hours of video footage. We evaluate the HTK toolkit, a state-of-the-art speech recognition engine, in combination with multiple video feature descriptors, for both the recognition of cooking activities (e.g., making pancakes) as well as the semantic parsing of videos into action units (e.g., cracking eggs). Our results demonstrate the benefits of structured temporal generative approaches over existing discriminative approaches in coping with the complexity of human daily life activities.

1. Introduction

Human activity recognition has been the subject of extensive research. In recent years, popular topics have emerged from video monitoring and surveillance to activity detection and behavioral analysis. One of the main challenges associated with the recognition of purposeful human actions is their inherent variability. One good measure of the productivity of the field is the sheer number and variety of datasets that have been produced by researchers in the last few years: A recent survey [3] lists no less than twenty-three common benchmarks. However, these rapid advances also demand an increased level of scrutiny.

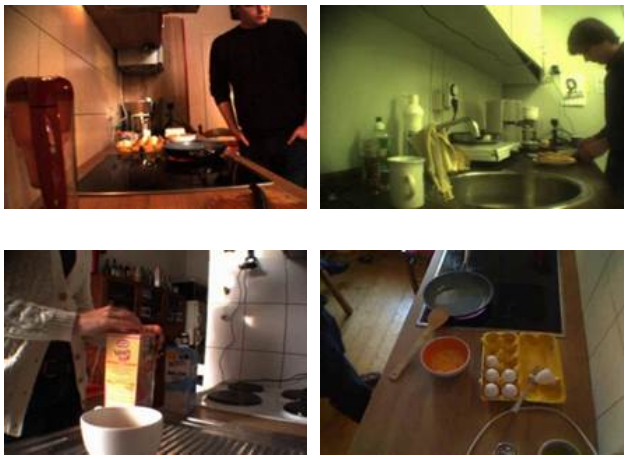


Figure 1. Sample images from the proposed breakfast dataset.

Rohrbach *et al*. describe several common shortcomings of present datasets [17]: Existing datasets tend to focus on coarse grained activities, they often exhibit artificially high inter-class variability, and videos tend to be (manually) pre-segmented. We recognize these shortcomings and offer additional points of consideration. First, many human activities are goal-directed, sequential, and require the use of tools. This is only partly reflected in current datasets. Second, most datasets have been collected in highly controlled setups using scripted actions recorded from one single environment under a limited set of viewpoints. Overall, this limits greatly the ability of these systems to generalize to more real-life environments, actors or objects.

To overcome these issues, we collected a new dataset, composed of video recordings of unscripted actions in completely natural settings. This allows the formulation of a number of novel research questions and, as we will show, this novel dataset addresses several of the points mentioned above. The video dataset consists of 10 cooking activities

*Part of this work was done at the lab for Computer Vision for Human-Computer Interaction at the Karlsruhe Institute of Technology (http://cvhci.anthropomatik.kit.edu/).

performed by 52 different actors in multiple kitchen locations (see Fig. 1 for sample frames). One focus of the proposed dataset is on the levels of granularity of human activities.

Research in cognitive psychology has shown that human participants do not perceive events as flat, but instead as hierarchical structures: It is possible to segment a continuous stream into distinct meaningful events at multiple levels of granularity starting from individual motor commands to composite sub-goals. In addition, studies, *e.g.* by Hard *et al.* [8], have shown that the amount of information contained in a video sequence around breakpoints depends on the action granularity of the corresponding breakpoints, with breakpoints associated with coarser action units carrying more information than breakpoints associated with finer units. Interestingly, the concept of granularity has received little attention in the context of video annotations [17, 19] despite their high significance for visual cognition.

To overcome the natural challenges that may arise when modeling human activities, we propose to model action recognition as a structured temporal process. We use an open source automated speech recognition engine, the Hidden Markov Model Tool Kit (HTK) [25]. Concepts from speech recognition can be naturally transposed to activity recognition: Coarsely labeled action units, modeled by Hidden Markov Models (HMMs), much like words in speech, form the building blocks for longer activity sequences. Units are combined into sequences using an action grammar. Beside action recognition, this approach enables the semantic parsing as well as the segmentation of videos at the level of single frames. An overview of this approach is shown in Fig. 2.

The system is evaluated both on the proposed dataset as well as a standard benchmark dataset, the Activities of Daily Living (ADL) [15]. Our results show the benefits of structured temporal approaches compared to standard discriminative approaches in coping with increasingly complex human behaviors.

2. Related work

Action recognition has benefited from an overwhelming variety of techniques. Specifically, an approach that has recently been made popular is the re-purposing of tools used in language and speech research. An early approach to model activities via hierarchical models (which also uses an early version of HTK) was proposed by Ivanov and Bobick [9]. The authors use trajectories gained from different visual tracking systems to evaluate different human activities from basic hand gestures to music conducting to understanding patterns of traffic in a parking lot. In recent years, sequence modeling tools have received an increasing amount of attention, as shown, for instance, in the survey by Weinland *et al.* [24], in part because of their ability to

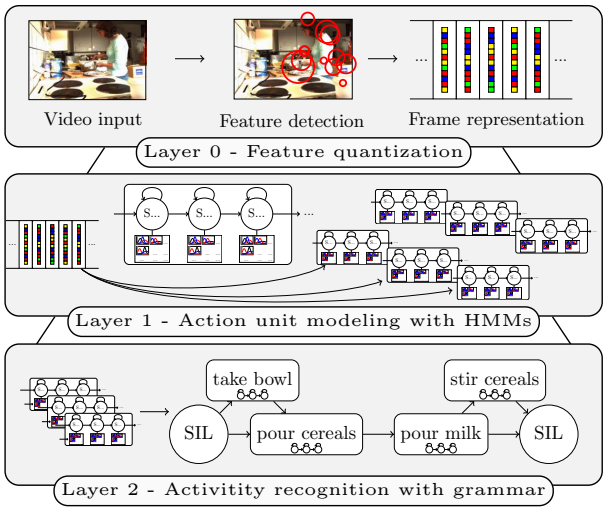


Figure 2. Overview of the proposed hierarchical approach for the recognition and parsing of human activities using the open source speech recognition framework HTK.

model the temporal structure of actions at multiple levels of granularity. HMMs are a particularly popular framework as they have been shown to work well not only for the recognition of events but also for the parsing and segmentation of videos [10] with applications ranging from sign language understanding[6, 14] to the evaluation of motor skills including the training of surgeons [26].

In the context of the recognition of human actions in video, Chen and Aggarval [5] use the output of an SVM to classify complete activities with HMMs, reaching a recognition accuracy of 90.9% on the KTH dataset. Other attempts also demonstrated the capability of HMMs [1, 22], but so far, HMMs have not reached state-of-the-art accuracy and lag behind discriminative approaches.

One limitation for scaling up HMMs to more complex scenarios is the use of Gaussian Mixture Models (GMMs), which require a dense low dimensional input representation in order to work well. As a result, behavioral analysis with HMMs is often done with motion capture data or other sensors [10] or, in the case of video-based action recognition, with object, hand and head trajectories [6, 26]. This has typically forced researchers working with HMMs to work in controlled environments with restrictive setups.

Another challenge associated with the use of HMMs for action recognition is that they require large amounts of training data. In speech recognition, where HMMs proved to be most successful, even an outdated corpus like TIDIGITS [12] for speech-based recognition of connected digits comprises 326 speakers uttering 77 digit sequences each, resulting in nearly 9,000 voice samples. More recent corpora like the Quearo Corpus have now reached over one million words [21]. These corpora are orders of magnitude larger than present video datasets for action recognition.

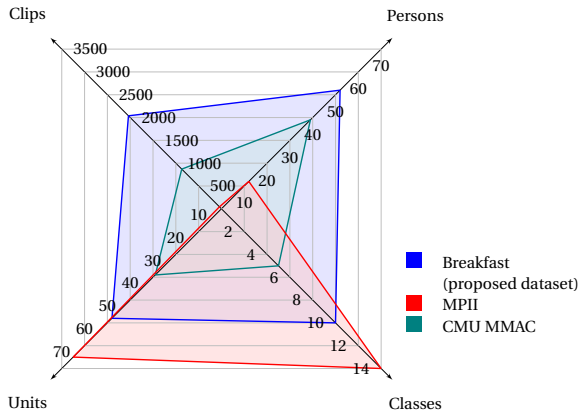


Figure 3. Comparison of different action recognition datasets based on the number of subjects, classes, action units and clips.

Another limitation of existing datasets is that many lack annotations for basic atomic entities within individual sequences, which are needed to train temporal models like HMMs. However, the annotation of those atomic entities is extremely time consuming and, therefore, it is only available for a small number of datasets. Some of the most recent action recognition datasets with this level of annotation include the CMU-MMAC dataset [19], the ICPR-KSCGR dataset [18], the GTEA Gaze+Dataset [7], the MPII Cooking dataset [17] as well as the 50 Salads dataset [20]. A comparison between the largest of these datasets and the proposed one is shown on Fig. 3.

3. Breakfast dataset

Data collection and pre-processing: The proposed dataset is to date one of the largest fully annotated datasets available. Overall, we recorded 52 unique participants, each conducting 10 distinct cooking activities captured in 18 different kitchens[1]. One of the main motivations for the proposed recording setup "in the wild" as opposed to a single controlled lab environment [17, 19], is for the dataset to more closely reflect real-world conditions as it pertains to the monitoring and analysis of daily activities.

The number of cameras used varied from location to location ($n = 3 - 5$). The cameras were uncalibrated and the position of the cameras changes based on the location. Overall we recorded ∼77 hours of video (> 4 million frames). The cameras used were webcams, standard industry cameras (Prosilica GE680C) as well as a stereo camera (BumbleBee®, Pointgrey, Inc). To balance out viewpoints, we also mirrored videos recorded from laterally-positioned cameras. To reduce the overall amount of data, all videos were down-sampled to a resolution of 320×240 pixels with a frame rate of 15 fps.

[1] http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset

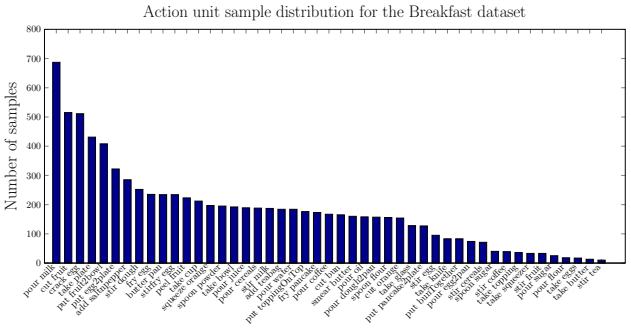


Figure 4. Number of samples per action unit in the dataset.

Cooking activities included the preparation of coffee (n=200 samples), orange juice (n=187), chocolate milk (n=224), tea (n=223), a bowl of cereals (n=214), fried eggs (n=198), pancakes (n=173), a fruit salad (n=185), a sandwich (n=197) and scrambled eggs (n=188). This set of goal-directed activities that people commonly perform in kitchens include both very distinct (*e.g.* tea and fried egg preparation) as well as very similar activities (*e.g.* fried egg vs. scrambled egg preparation) to allow for a comprehensive assessment of the recognition system. Unlike most existing datasets, the actor performance here was completely unscripted, unrehearsed and undirected. The actors were only handed a recipe and were instructed to prepare the corresponding food item. The resulting activities are thus highly variable both in terms of the choice of individual action units by the actors as well as in their relative ordering.

Data annotation: We asked two sets of annotators to manually label the corresponding videos at two different levels of granularity: One group of three annotators was asked to annotate action units at a coarse level, like 'pour milk' or 'take plate'. Another group of fifteen annotators was asked to provide annotations at a finer level of granularity by decomposing coarse actions such as 'pour milk' into finer chunks such as 'grab milk' → 'twist cap' → 'open cap.' In the present work, we only consider the coarse level of annotations. Analysis using finer action units will be published elsewhere. Overall we identified 48 different coarse action units with 11,267 samples in total including ∼3,600 'silence' samples (see Fig. 4 for labels used and the corresponding number of samples and Tab. 1 for the breakdown of action units for individual activities). The order the units can appear in is defined by a grammar (Fig. 5) built in a bottom-up manner using the original labels.

Data splits: For evaluation purpose, we organized the 52 participants in four groups, and permuted each of these four groups as splits for training and test. Because of the very large size of the dataset we found that using larger and/or more numerous splits became rapidly unpractical for any kind of extensive empirical evaluation.

Coffee	take cup - pour coffee - pour milk - pour sugar - spoon sugar - stir coffee
(Chocolate) Milk	take cup - spoon powder - pour milk - stir milk
Juice	take squeezer - take glass - take plate - take knife - cut orange - squeeze orange - pour juice
Tea	take cup - add teabag - pour water - spoon sugar - pour sugar - stir tea
Cereals	take bowl - pour cereals - pour milk - stir cereals
Fried Egg	pour oil - butter pan - take egg - crack egg - fry egg - take plate - add salt and pepper - put egg onto plate
Pancakes	take bowl - crack egg - spoon flour - pour flour - pour milk - stir dough - pour oil - butter pan - pour dough into pan - fry pancake - take plate - put pancake onto plate
(Fruit) Salad	take plate - take knife - peel fruit - cut fruit - take bowl - put fruit to bowl - stir fruit
Sandwich	take plate - take knife - cut bun - take butter - smear butter - take topping - add topping - put bun together
Scrambled Egg	pour oil - butter pan - take bowl - crack egg - stir egg - pour egg into pan - stir fry egg - add salt and pepper - take plate - put egg onto plate

Table 1. Breakdown of action units for individual activities. The ordering of action units for each activity is defined separately using a grammar.

4. Modeling actions as speech

As Chen and Aggarval pointed out, there is an inherent similarity between the temporal structure of actions and speech [5]. Fine-level actions can be combined into coarser action units much like phonemes for words in speech recognition. Longer activities can be further constructed using a grammar over action units. This analogy between action and speech processing suggests that the tools and techniques from automatic speech recognition could, in principle, be applicable to action video data. Here we use the HTK [25], an open source toolbox mainly dedicated to speech recognition and adapt it for the recognition and parsing of human activities. HTK provides the tools for evaluation and decoding of sequences and also supports data formats and vocabularies that are not based on audio input.

4.1. Modeling action units

We start modeling human activities at the level of action units, corresponding to the coarser labels of the dataset. Each action unit is modeled by a single HMM following the modeling of phonemes as described in [25]. To initialize the HMMs, training units are equally divided by the predefined number of states. The initial Gaussian components are computed from the training samples and initial transition probabilities set to constant values for self-loops as well as transitions to the next state. Special cases, namely the transitions from the start and end state, are treated separately by setting the start state transition to 1 and the end state transition to 0. The parameters are optimized separately for each HMM using Baum-Welch re-estimation. For recognition and decoding of HMMs, HTK features the Viterbi algorithm to compute the best path at each time step given an input sequence. This is done by summing up log transition probabilities and log output probabilities of the most probable states. For specific details of the training and decoding of HMMs in HTK, we refer to [25].

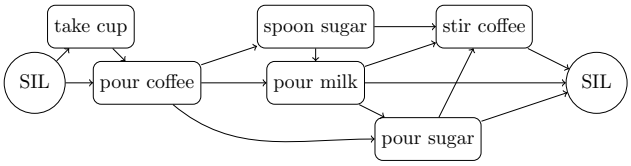


Figure 5. Possible transitions between action units for the activity "preparing coffee".

4.2. Modeling of complex activities

At the higher level, activities are modeled as sequences of action units using a context-free grammar. This two-stage approach has several benefits: First, breaking down the complete tasks into smaller units allows not only the recognition of the activity as a whole, but also its parsing at the level of action units. The result is a unit-level representation and segmentation of an input sequence. Second, the bottom-up construction of tasks from single independent units allows a higher flexibility when dealing with unknown combinations of units up to the recognition of new, unseen activities.

Following the HTK convention, we use a grammar specified in the Extended Backus Naur Form (EBNF). The recognition of sequences is based on the token passing concept for connected speech recognition [25], augmenting the partial log probability with word link records, or in the present case, unit link records describing the transition from one unit to the next. The Viterbi algorithm is used to compute the most probable sequence. For any given time point, the link records can be traced back to get the current most probable path. This corresponds to the most probable combination of units, and the position of the related unit boundaries, that is essentially the segmentation of the sequence.

5. Evaluation

5.1. Feature representation

To evaluate the performance of the proposed approach, we consider two popular visual descriptors, the HOGHOF proposed by Laptev [11] and Trajectons as described by Matikainen *et al.* [13]. Here we apply a bag-of-words approach [11, 23]: To build the codebooks, 100k features were randomly sampled from the training set and clustered into K clusters using the K-means algorithm with $K = \{30, 50, 100, 200\}$. The resulting cluster centers were used to build histograms over a sliding temporal window by hard assignment of each feature to its cluster center.

5.2. Recognition with HTK

For training, an initial HMM is initialized for each action unit specifying the number of stages, the topology (in this case forward left-to-right), the initial transition probabilities (0.6 - self, 0.4 - next) and the number of Gaussians per state. The number of stages was determined by cross validation using $n = [3, 5, 10, 15]$ fixed stages. We also tested an adaptive number of stages depending on the mean frame length of the related unit and linear scaling (factor 10). Overall linear scaling showed the best results for all cases. Additionally, the number of Gaussians was evaluated with $m = [1, 2, 3, 5, 10]$ Gaussians. Here, the modeling with a single Gaussian distribution outperformed the other models. Given that an adaptive number of stages and a single Gaussian consistently gave higher accuracy, these were set as default parameters in all remaining experiments.

To validate the proposed approach, we applied the system to an independent benchmark dataset, the Activities of Daily Living (ADL) [15], achieving a recognition accuracy of 77.3 % for HOGHOF and 74.0% for Trajectons. As the dataset is rather small for HMMs (only 5 test persons and 150 clips in total), the results do not quiet outperform the current state-of-the-art (82.7% [16]). Nevertheless, the recognition still outperforms the originally reported 69% accuracy for Laptev's HOGHOF [15] and shows a comparable level of accuracy with other approaches thus demonstrating the potential of the proposed framework.

5.3. Accuracy measures for unit recognition

To compare the output of the unit recognition module to the ground-truth we consider two different types of errors.

First, as a measure for the correct detection of a unit within a sequence, we report the unit accuracy based on the concept of word accuracy in speech. A direct comparison between recognized units vs. ground-truth is usually not possible, as the number of recognized units does not necessarily match the number of units in the reference sequences. Therefore, we first align the recognized sequence to the reference sequence by dynamic time warping (DTW)

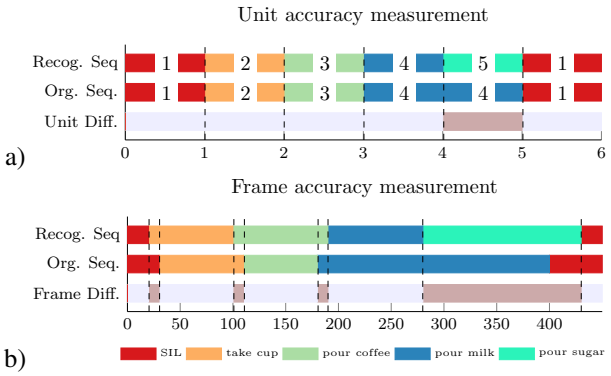


Figure 6. Evaluation methods used: Comparison between **a)** unit-based vs. **b)** frame-based performance measurement.

as an evaluation preprocessing step. After alignment, as shown in Fig. 6 a), three different types of errors, insertions, deletions and misclassifications can occur. Different from speech, where errors are reported in relation to the number of units in the original sequence, we consider a more relaxed measurement by just considering all units after DTW without distinguishing between insertions, deletions or misclassifications. Second, we report the frame-based accuracy as a measure of the quality of segmentation. For frame-based accuracy, we measure the number of frames labeled incorrectly for the related sequence as shown in Fig. 6 b).

5.4. Evaluation of Breakfast dataset with HTK

Sequence recognition: First, we evaluate the overall sequence recognition accuracy for the 10-class activity recognition problem. We tested the proposed system with different codebook sizes $K = \{30, 50, 100, 200\}$ for HOGHOF and Trajectons (see Tab. 2). We found that HOGHOF perform best at 38.46% outperforming Trajectons with an accuracy of 28.68%.

Looking at the confusion matrix for HOGHOF shown on Fig. 7, it shows that related activities like the preparation of drinks vs. food tend to be more often confused. We see two reasons for that: First, related activities share a higher number of similar action units compared to unrelated activities. Second, the combination of HMMs and grammar does encode implicitly the possible length of a sequence. Much like HMMs require a sufficient number of frames for each state, grammars also define a minimum number of HMMs. It is unlikely that very long activities, like the preparation of pancakes, get mixed up with very short ones like preparing coffee. The only exception is the preparation of cereals. As Fig. 7 shows, the "cereals" sequences tend to be confused with the activities involving drinking more than with food-related activities. We attribute this to the fact that, even though it does not fit the literal grouping, this activity shares more units with drink-related activities like stirring and pouring than with food-related activities. Also

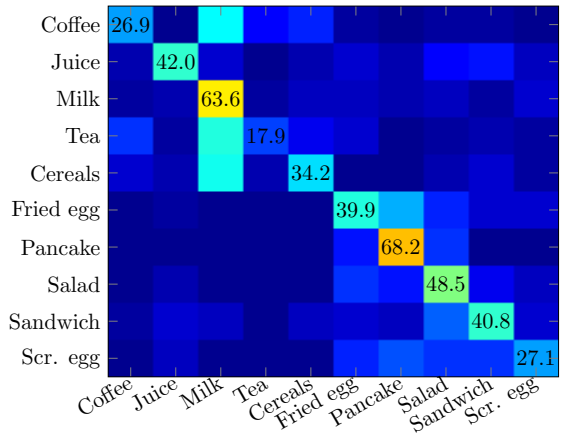


Figure 7. Confusion matrix for activity recognition (HOGHOF). Closely related activities such as drinking or eating tend to be more easily confused.

Sequence recognition for 10 activity classes				
	c30	c50	c100	c200
HOGHOF	38.5%	**40.5%**	38.9%	39.4%
Trajectons	**28.7%**	27.1%	26.7%	27.1%

Table 2. Activity recognition performance using HTK for different codebook sizes.

the mean duration of this activity tends to be more similar to drink-related than to food-related activities.

In addition, there also appears to be a trend towards better recognizing longer activities. When looking at the best activity of the food and drink group, both the preparation of pancakes and chocolate milk also happened to be the longest activities of their related groups. In general, drink-related activities tend to be confused with the "chocolate milk," which is also the longest of all. In both cases, the drink and the food preparation, longer activities are favored over short ones.

Unit recognition: At the level of action units, we evaluate the performance of the unit accuracy and frame-based segmentation as described in Sec. 5.3. Obligatory leading and trailing silence units at the beginning and end of each sequence are ignored as they are predefined by the grammar and thus, can not be seen as result of the recognition process. As shown in Tab. 3, for the case of the HOGHOF-based recognition, 31.8% of all units and 28.8% of all frames were correctly recognized.

Considering that this result also includes misclassified activities, we evaluated the unit accuracy for correctly classified acticities leading to an overall unit accuracy of 65.7% for HOGHOF and 64.0% for Trajectons and a frame-based recognition rate of 58.5% for HOGHOF and 56.8% for Trajectons.

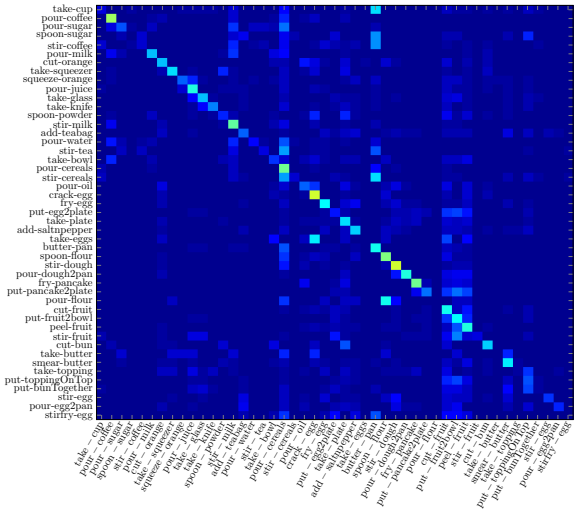


Figure 8. Confusion matrix for unit recognition (HOGHOF).

Unit accuracy results for 48 unit classes				
	c30	c50	c100	c200
HOGHOF	30.4%	**31.8%**	31.3%	31.7%
Trajectons	**23.0%**	21.6%	21.8%	21.5%

Frame-based accuracy results for 48 unit classes				
	c30	c50	c100	c200
HOGHOF	**28.8%**	**28.8%**	26.6%	26.6%
Trajectons	**24.5%**	24.2%	22.5%	**24.5%**

Table 3. Unit and frame-based recognition accuracy for all 48 action units.

5.5. Grammar-based recognition

As the proposed system has a hierarchical structure, the grammar that guides the overall recognition plays an important role for the final sequence parsing and frame-based recognition. To separate the influence of the grammar from the simple HMM-based parsing, we evaluate the proposed system by replacing the grammar describing complete sequences by a flat grammar, allowing transitions from each HMM to any other. To limit the length of the sequences, the minimum number of units that needs to be used is set to 2 and the maximum number is set to 15. Overall, one can see that the unit accuracy is at best 12.5% for HOGHOF and 9.2% for Trajectons and frame-based accuracy is at best 12.4% for HOGHOF and 13.2% for Trajectons, but still outperforms frame-based unit classification by SVM (see sec. 5.6). This may be due to the fact that even a flat grammar is still more constrained by giving a minimum and maximum number of possible unit transitions and that HMMs encode the temporal evolution of different units, which is not the case for a simple frame based SVM classification.

Unit accuracy for 48 units w/o grammar				
	c30	c50	c100	c200
HOGHOF	10.9%	11.9%	12.3%	**12.5%**
Trajectons	8.4%	8.8%	8.1%	**9.2%**

Frame-based accuracy for 48 units w/o grammar				
	c30	c50	c100	c200
HOGHOF	12.1%	12.1%	**12.4%**	12.0%
Trajectons	11.6%	12.0%	11.7%	**13.2%**

Table 4. Unit and frame-based recognition accuracy for HTK without grammar.

Results of SVM and RF classification for 10 activity classes					
		c30	c50	c100	c200
SVM	HOGHOF	25.1%	**26.0%**	21.5%	21.0%
	Trajectons	22.3%	21.7%	23.2%	**26.0%**
RF	HOGHOF	22.7%	**24.0%**	20.9%	22.7%
	Trajectons	18.1%	21.6%	22.4%	**25.1%**

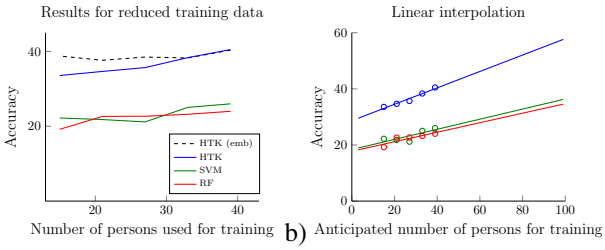
Table 5. Recognition accuracy for SVM and RFs for overall activities.

5.6. Comparison with discriminative classification

To evaluate the performance of the proposed system compared to state-of-the-art action recognition as described in [11, 23], we consider a standard discriminative classification approach using SVMs with RBF kernels [2] as well as random forests (RFs). To estimate the best parameters for C and γ, a five-fold cross-validation procedure was used together with an all-pair multi-class classification strategy. The Synthetic Minority Over-sampling Technique (SMOTE) [4] was used whenever needed (*i.e.*, for highly imbalance unit- and frame-based classification problems). To avoid oversampling with too many artificially generated data, the maximum number of samples per class was limited to 1,000 samples.

We classify the overall activity for each clip as well as the action unit for each frame. For activity recognition, discriminative classifiers were trained with histograms built from all features of the complete video clip. For the frame-based unit classification, histograms were built over a 10-frame sliding temporal window. Tab. 5 shows results for the discriminative models for activity recognition: the maximal accuracy was 26.0% for HOGHOF for $K = 50$ and Trajectons for $K = 200$, whereas recognition performance with HTK lies at 38.46% (+12.42%) for HOGHOF ($K = 50$) and 28.68% for Trajectons (+2.68%, $K = 30$). We also extended the codebook size for activity recognition to $K = 2,000$ words because this is closer to dictionary sizes used by other authors [11, 23] but accuracy did not improve (22.4% for HOGHOF and 24.0% for RFs).

Further, we also computed the unit accuracy per frame which is at best at 6.09% for HOGHOF and 6.33% for Trajectons. This is still a significant drop compared to the recognition with HTK with and without grammar and shows that a simple classification of time snippets is not sufficient given the complexity of the data.



a) b)

Figure 9. Activity recognition with a reduced number of training samples: A comparison between HTK, SVM and RFs (a) and extrapolation to 100 training samples (b).

5.7. Number of training samples needed

One of the main limitations of the system is the need for annotations at the level of individual action units, which is a long and tedious task for annotators.

Reduced training samples: On Fig. 9, we evaluate how the accuracy of HTK vs. discriminative approaches (SVM and RFs) vary as a function of the number of video samples used for training. We reduced the number of people used for training from 40 down to 15 (i.e., 5 people per split). This is the smallest training set that still includes all possible action units. It corresponds to 500 clips (i.e., 50 training samples per sequence).

As can be seen on Fig. 9 a), HTK outperforms discriminative methods, even with a reduced number of training samples. We also applied a linear fitting to the data as displayed in Fig. 9 b).

Embedded training and flat start: As a second option to reduce the labeling workload, we explored the capabilities of embedded training (see [25], chap. 8.5). For this bootstrapping procedure, only part of the training data needs to be labeled at the level of action units to initialize the HMMs in a first training run. For the remaining data, only the transcription of the occurring units is needed. The transcribed data is used for a refinement of the initialized HMMs data by embedded Baum-Welch re-estimation. Fig. 9 a) shows the results of the embedded training using the indicated number of persons for initialization and, in the second step, using the transcripts of all 39 persons in the training split to re-estimate the models.

In the case that no unit labels are available, this technique also allows a so called "flat start". Here, untrained HMMs are used as prototypes and the re-estimation is applied for an iterative segmentation and training based on transcribed data. The flat start training showed a sequence accuracy of 25.34%.

6. Conclusion

We described an effort to advance the field of human activity recognition by applying techniques borrowed from speech recognition. To evaluate the approach, we collected a novel activity dataset that is both objectively large and challenging. We trained a hierarchical model based on HMMs and a context-free grammar using the open source HTK speech recognition toolbox. We evaluated the approach in the context of both video parsing and classification for ten different kitchen activities. We demonstrated the potential of temporally structured, generative models for activity recognition, reiterating the need for copious amounts of video data for these models to perform well. We hope that this initial attempt and the resulting video dataset will spur further research towards a large scale recognition of complex, goal-directed, daily-life activities.

Acknowledgments

This work was supported by ONR grant (N000141110743) and NSF early career award (IIS-1252951) to TS. Additional support was provided by the Robert J. and Nancy D. Carney Fund for Scientific Innovation and the Center for Computation and Visualization (CCV). HK was funded by the Quaero Programme and OSEO, the French State agency for innovation.

References

[1] A. Antonucci, R. de Rosa, and A. Giusti. Action recognition by imprecise hidden markov models. In *IPCV*, 2011. 2

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011. 7

[3] J. M. Chaquet, E. J. Carmona, and A. Fernndez-Caballero. A survey of video datasets for human action and activity recognition. *CVIU*, 117(6):633 – 659, 2013. 1

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002. 7

[5] C. Chen and J. Aggarwal. Modeling human activities as speech. In *CVPR*, 2011. 2, 4

[6] P. Dreuw, J. Forster, and H. Ney. Tracking benchmark databases for video-based sign language recognition. In *ECCV Int. Workshop on Sign, Gesture, and Activity*, 2010. 2

[7] A. Fathi, Y. Li, and J. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 3

[8] B. M. Hard, G. Recchia, and B. Tversky. The shape of action. *Journal of experimental psychology: General*, 140(4):586–604, Nov. 2011. 2

[9] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, 2000. 2

[10] D. Kulic and Y. Nakamura. Incremental learning of human behaviors using hierarchical hidden markov models. In *IROS*, 2010. 2

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 5, 7

[12] R. G. Leonard and G. Doddington. TIDIGITS, Texas Instruments, Inc., 1993. http://catalog.ldc.upenn.edu/LDC93S10. 2

[13] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *ECCV*, 2010. 5

[14] S. Mehdi and Y. Khan. Sign language recognition using sensor gloves. In *Int. Conf. on Neural Information Processing*, 2002. 2

[15] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009. 2, 5

[16] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *ECCV*, 2010. 5

[17] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012. The dataset and relevant code is available at http://www.d2.mpi-inf.mpg.de/mpii-cooking. 1, 2, 3

[18] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern. Kitchen scene context based gesture recognition: A contest in ICPR2012. In *Advances in Depth Image Analysis and Applications*, LNCS 7854, 2013. 3

[19] E. H. Spriggs, F. De la Torre Frade, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Workshop on Egocentric Vision, CVPR*, June 2009. 2, 3

[20] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM Int. Conf. on Pervasive and Ubiquitous Computing*, 2013. 3

[21] S. Stüker, K. Kilgour, and F. Kraft. Quaero 2010 speech-to-text evaluation systems. In *High Performance Computing in Science and Engineering*. 2012. 2

[22] R. Vezzani, D. Baltieri, and R. Cucchiara. HMM based action recognition with projection histogram features. In *ICPR*, 2010. 2

[23] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 5, 7

[24] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *CVIU*, 115(2):224 – 241, 2011. 2

[25] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. The HTK book, version 3.4. Technical report, Cambridge University Engineering Department, Cambridge, UK, 2006. 2, 4, 7

[26] Q. Zhang and B. Li. Relative hidden markov models for evaluating motion skills. In *CVPR*, 2013. 2