Discovering Primitive Action Categories by Leveraging Relevant Visual Context

Kris M. Kitani, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto

Discovering Primitive Action Categories by Leveraging Relevant Visual Context

Kris M. Kitani, Takahiro Okabe, Yoichi Sato
The University of Tokyo
Institute of Industrial Science
Tokyo 158-8505, Japan
{kitani, takahiro, ysato}@iis.u-tokyo.ac.jp

Akihiro Sugimoto
National Institute of Informatics
Tokyo 101-8430, Japan
sugimoto@nii.ac.jp

Abstract

Under the bag-of-features framework we aim to learn primitive action categories from video without supervision by leveraging relevant visual context in addition to motion features. We define visual context as the appearance of the entire scene including the actor, related objects and relevant background features. To leverage visual context along with motion features, we learn a bi-modal latent variable model to discover action categories without supervision. Our experiments show that the combination of relevant visual context and motion features improves the performance of action discovery. Furthermore, we show that our method is able to leverage relevant visual features for action discovery despite the presence of irrelevant background objects.

1. Introduction

We present a novel framework for the unsupervised learning of primitive human actions from a video corpus by leveraging *visual context*, where we define visual context as the appearance of the entire scene including the actor, related objects and relevant background features. Since actions can be understood at various temporal resolutions, we focus on the discovery of what we call *primitive actions*. Primitive actions are human actions that can be defined over a very short period of time (a few seconds). For example, grabbing a cup, typing on a keyboard or flipping the page of a book can be recognized within a few seconds of observing the action. Learning primitive actions are important because they are the basic building blocks of many high-level activities [17, 9, 12].

Supervised learning techniques using such models as HMMs [28, 11], Bayesian classifiers [26] and temporal dynamics [27] have been successful in describing primitive actions but need labeled data or a considerable amount of prior knowledge. While most of these past works have used supervised approaches to learn primitive actions, there have also been several recent works focused on the unsupervised learning of actions.

An approach of growing interest for unsupervised topic discovery is the use of generative latent variable models (mixture models [21], pLSA [10], LDA [2], HDP [29]) based on the bag-of-words paradigm. Most of these methods were originally applied to text processing to learn topics from text in documents. Given the document analogy, a typical language model factorizes the observed documents (corpus items) and words (features) to discover a distribution over a set of hidden topics.

Niebles [20] proposed the application of a generative model to video to learn action categories (topics) from a bag-of-features. They used exactly the same framework as [10] by simply replacing document indices with video indices, and words with spatial-temporal (ST) volumes. Their approach showed that similar to text, the local features of an action can be treated as though they were *exchangeable* (an action can be treated as a bag of uncorrelated features) to learn action categories. However, the conceptual problem with a straightforward use of a language model for action discovery is that the models are uni-modal (e.g. use only words or ST features).

We know from experience that many actions are composed of motions and visual appearance. For example, the hands of a person playing a piano and typing on a keyboard might have very similar motions but can easily be differentiated using the visual context of a piano or a keyboard. In fact, findings from neural science show that actions are mentally perceived as a mix of motions and visual features of present objects [6]. In the light of this fact, many previous approaches to action discovery are limited by the fact that they only consider actions defined strictly by motion. For example, the ST volumes used to describe dynamic human actions in [20] are robust against static spatial features but are also insensitive to visual features produced by related static objects in the scene. Similarly, the approach proposed by Wang [30] to automatically classify actions of cars and pedestrians at an intersection uses the change in pixel

Figure 1. Leveraging visual features for action recognition: Relevant visual features (green) induced by using the telephone and irrelevant features (purple) produced by unrelated background objects.

intensity between two frames as the input feature, making their system robust to the varying color and shape of automobiles and pedestrians. However, to apply their current system to actions that involve interactions between actors and objects will most likely necessitate the incorporation of more complex motion features and appearance features to improve performance.

While the joint use of appearance and motion to describe action is not entirely new, our work differs from previous work in that we leverage appearance information without the use *a priori* information about the shape or color of actors or objects in the scene. Work using the appearance of related objects to recognize actions has depended on *a priori* knowledge of the appearance of related objects [8] or pre-defined object categories [18].

Work leveraging the appearance of the actor, such as Fanti [7] and Niebles [19], has proposed modified generative models that model the human body and account for both the appearance and motion of body parts. However, explicitly modeling the human body comes at the cost of losing the ability to apply the model to other types of actors. The more important distinction with our work however, is that the use of visual information was limited to pre-defined body parts and therefore could not explicitly take into account other relevant visual information possibly generated by co-occurring objects or scenic context.

In this work, we present a robust framework for primitive action discovery by leveraging both motion and relevant visual context without the use of *a priori* information (e.g. an explicit shape model or pre-defined object categories). Our experiments show that our method properly leverages relevant visual appearance and is robust against irrelevant visual features (Figure 1) when learning action categories.

2 Proposed method

Our goal is to learn the primitive action categories that occur within a video corpus. First we extract temporal features and spatial features from each video segment, under the assumption that actions are defined by both temporal motion and visual context (Section 2.1). Then we describe a dimension reduction process to create a codebook for each feature type (Section 2.2). Finally we describe the bi-modal latent variable model that uses the histograms produced by motion and visual context to learn the latent action categories (Section 2.3).

2.1. Extracting spatial and temporal features

Here we begin by explaining how spatial (visual) features and temporal (motion) features are extracted. For each frame in the training corpus, we extract a sparse set of spatial features by finding SIFT keypoints [16]. These keypoints are then represented with a normalized 128-d SIFT descriptor. Other combinations of keypoint detectors and descriptors can be used as well.

Using the same temporal gradient descriptor as [3], a set of temporal features are extracted from the video frames by extracting a $7 \times 7 \times 4$ (a 7×7 spatial window over 4 frames) spatio-temporal volume for pixels detected as a good feature to track [25]. Each element of the volume contains the temporal gradient magnitude. The descriptor is a normalized 196-d vector containing the elements of the volume. More complex temporal keypoints can also be used, such as spatio-temporal cuboids [5] or space-time interest points [13].

2.2 Feature clustering

For each mode (temporal and spatial) we first implement the standard k-means clustering algorithm to roughly cluster the features and produce the codebook $\mathbf{C}$. A feature histogram $\mathbf{v}_d$ for each video segment d in the corpus $\mathbf{d}$ is created by assigning each extracted feature to the closest cluster. This process yields a set of histograms $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_m)$, where $m = |\mathbf{d}|$. We then further reduce the dimensionality of the histograms using a more holistic approach, namely, non-negative matrix factorization (NMF) [14] which projects the set of histograms $\mathbf{V}$ onto a lower dimensional non-negative subspace $\mathbf{H}$. Since we assume that actions are additive (actions only produce features and never delete features), we use NMF over other projection techniques like PCA or ICA because NMF decomposes the data $\mathbf{V}$ as an additive (non-negative) combination of a lower dimensional basis subspace. It is interesting to note that both

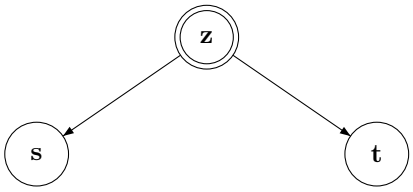

Figure 2. Bi-modal latent variable model defined by the latent topic $\mathbf{z}$, spatial features $\mathbf{s}$ and temporal features $\mathbf{t}$.

pLSA and NMF have been shown to be instances of multinomial PCA [4]. In fact in our framework, the projected dimensions of NMF r and the number of hidden categories q are set to be equivalent and as a consequence this second stage can also be interpreted to be the pre-discovery of hidden actions categories for each independent modality of input.

Formally, NMF decomposes the $k \times m$ histogram matrix $\mathbf{V}$ (each column $\mathbf{v}_i$ is a k dimensional histogram of features for the i-th video) into a $k \times r$ basis matrix $\mathbf{W}$ and the $r \times m$ encoding matrix $\mathbf{H}$,

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

We use the projected gradient method [15] to factorize $\mathbf{V}$ and the resulting columns of the encoding matrix $\mathbf{H}$ contain the reduced (encoded) version of $\mathbf{V}$.

As with k-means clustering, we run NMF twice independently, once for spatial features and once for temporal features, by projecting the spatial and temporal descriptor histogram matrices $\mathbf{V}^s$ and $\mathbf{V}^t$ onto the reduced dimensional spaces $\mathbf{H}^s$ and $\mathbf{H}^t$, respectively. That is, NMF maps a histogram $\mathbf{v}_i$ from each video segment to a reduced dimensional histogram $\mathbf{h}_i$. As a result, the values of $\mathbf{H}^s$ and $\mathbf{H}^t$ yield an approximation to the term-by-document frequency matrix. The term-by-document frequency matrix is commonly used as the input for learning language models, where each element $n(w, d)$ of the matrix represents the number of times a feature w was observed in the corpus item d.

2.3 Merging motion and visual context via the action model

2.3.1 Parameter learning

Under the framework of Bayesian networks, the joint probability of a set of random variables can be simplified by defining the conditional independence between variables. In our action model (Fig. 2), we assume that temporal features $\mathbf{t}$ are conditionally independent of spatial features $\mathbf{s}$ given latent action categories $\mathbf{z}$. That is, we propose a bi-modal expansion of the standard mixture of uni-grams model [21] and define the probability of a video segment $d \in \mathbf{d}$ as below.

$$p(d) = \sum_z p(d|z)p(z) \tag{2}$$

$$p(d|z) \propto \prod_{s \in d} p(s|z) \prod_{t \in d} p(t|z) \tag{3}$$

$$= \prod_s p(s|z)^{n(s,d)} \prod_t p(t|z)^{n(t,d)} \tag{4}$$

Based on the conditional independence of a spatial feature s and a temporal feature t given the latent topic z, the conditional probability of a video segment can be computed as the product of the conditional probabilities of all spatial and temporal features in the video segment. The term $n(s, d)$ represents the number of times a spatial feature s has occurred in a video segment d. The term $n(t, d)$ is interpreted similarly for temporal features.

To learn the parameters of the bi-modal mixture model, we desire to find values for the parameters $p(s|z)$, $p(t|z)$ and $p(z)$, such that the log-likelihood of the entire video corpus $\mathbf{d}$ is maximized.

$$\log p(\mathbf{d}) = \sum_d \log \sum_z p(d|z)p(z) \tag{5}$$

We implement the expectation maximization (EM) algorithm to find a locally optimal set of parameters. Since it can be shown the lower bound of the data likelihood can be maximized by equivalently maximizing the expectation of the complete log-likelihood $E[\mathcal{L}^c]$ over the posterior $p(z|d)$, we maximize the following function.

$$E[\mathcal{L}^c] = \sum_d \sum_z p(z|d) \log p(d|z)p(z) \tag{6}$$

In the expectation step, the posterior of the latent variable is computed using Bayes' rule.

$$p(z|d) = \frac{p(d|z)p(z)}{\sum_{z'} p(d|z')p(z')} \tag{7}$$

By solving for the maxima of the Lagrangian function, which is composed of the complete data log-likelihood and standard conditions on the parameters (i.e. probabilities add to one), we obtain the following re-estimation equations for the maximization step. These re-estimation equations maximize the likelihood of the data given the current posterior.

$$\hat{p}(s|z) = \frac{\sum_d n(s, d)p(z|d)}{\sum_s \sum_d n(s, d)p(z|d)} \tag{8}$$

$$\hat{p}(t|z) = \frac{\sum_d n(t, d)p(z|d)}{\sum_t \sum_d n(t, d)p(z|d)} \tag{9}$$

$$\hat{p}(z) = \frac{\sum_d p(z|d)}{|\mathbf{d}|} \tag{10}$$

This process between the expectation step and the maximization step is repeated until the log-likelihood function converges at a maximum.

2.3.2 Inference and recognition

Although recognition is not the focus of this paper, once the parameters of the model have been learned, the naive Bayes model can also be used to recognize primitive actions. Specifically, given a test video segment d, a set of temporal and spatial features are extracted and binned to create a histogram of temporal and spatial features,$\mathbf{v}_d^t$ and $\mathbf{v}_d^s$ respectively. Then the histograms are projected onto the respective encoding spaces to obtain the encoding vectors $\mathbf{h}_d^t$ and $\mathbf{h}_d^s$ using the zeroed least-square solution [22]. By normalizing the vectors $\mathbf{h}_d^t$ and $\mathbf{h}_d^s$ we obtain the distribution over the features $\mathbf{t}_d$ and $\mathbf{s}_d$ for the test video segment. These distributions are then passed on as likelihood evidence for the naive Bayes model to infer the distribution over the hidden actions $p(z|d)$ using belief propagation [23].

3 Experiments

Publicly available datasets used for human action recognition, like the KTH dataset [24], have very little background variation (i.e. a wall or a field) and usually involve only actors with no interactive objects [5, 1, 31]. Here we present three new primitive action video datasets, created to include various backgrounds and interactive objects. These datasets are needed to show how our method is able to leverage relevant visual features along with motion information to effectively discover action categories.

3.1 Video datasets

3.1.1 Actions with objects corpus

The first motion and object corpus C_{OBJ} consists of eight different primitive actions that involve a related physical object. A list is given below:

1. Touch typing on keyboard
2. Beginner on keyboard
3. Dialing phone
4. Flipping pages of book
5. Skimming page of book with finger
6. Writing with pen on piece of paper
7. Sifting through stack of papers
8. Take cup

Each action video was spliced indiscriminately into three second intervals. Using the first five segments per action yielded a total of 40 video segments. Each video segment was 90 frames long and all videos were created at a resolution of 160×120. Notice that all of the actions are repetitive and sometimes involve the same object. Key frames from the corpus are given in Figure 3.

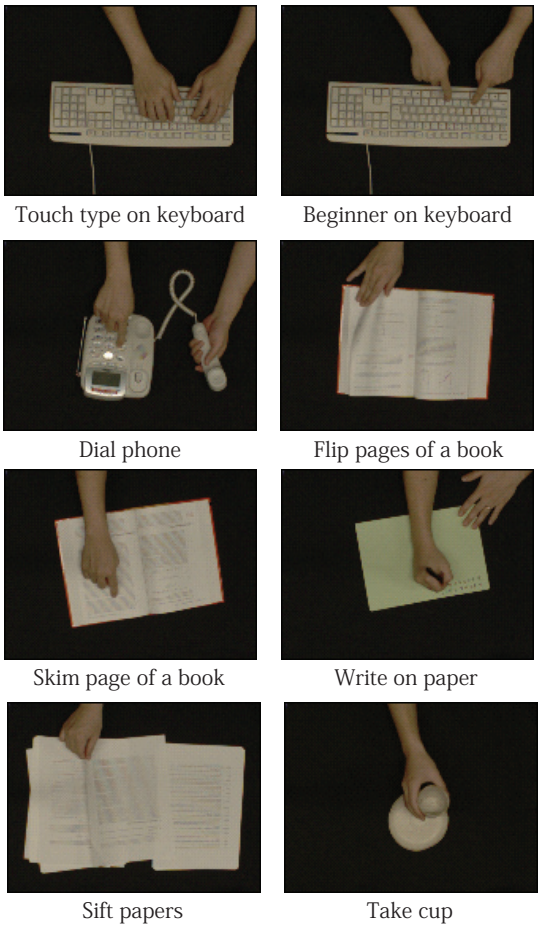

Figure 3. Key frames for the corpus C_{OBJ} with 8 desktop actions involving objects.

3.1.2 Actions with backgrounds corpus

The motion and background corpus C_{BG} consists of three different motions and three different visual scenes (backgrounds). The three motions are:

1. Take (vertical movement)
2. Wipe (horizontal movement)
3. Open (hand opens and closes in place)

The three background scenes are:

1. Board game scene
2. Tools scene
3. Cooking scene

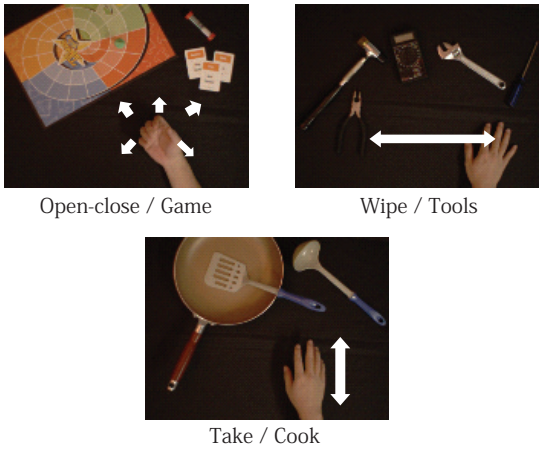

Figure 4. Examples from corpus C_{BG} with 9 actions including 3 different motions and 3 different background objects. Direction of motion is shown in white.

The combination of movement and scenes yields nine different actions. Five videos segments for each action resulted in a corpus of 45 videos. Each video segment was 90 frames long and all videos were created at a resolution of 160×120. Key frames of several combinations of motions and visual contexts are given in Figure 4.

3.1.3 Actions with objects and backgrounds corpus

The motion with objects and background corpus C_{BGOB} contains the same actions as the first corpus C_{OBJ} but also includes random backgrounds (Figure 5), which act as visual noise. The visual appearance of each video segment was varied by including different random objects in the background. This corpus used the same number of video segments and the same resolution as the first corpus C_{OBJ}.

3.2 Experimental setup and results

First, we performed one baseline experiment using the same uni-modal generative model as [20]. Then we performed three experiments using our new framework to show how leveraging visual context improves learning performance.

For each experiment the codebook was generated using k-means clustering, where $k = 1500$, the stop criteria was a cumulative displacement of cluster centers less than 1, the maximum number of iterations was 10 and distance was measured using the L_2-norm. We note that similar results were achieved for k within the range of 500 to 3000.

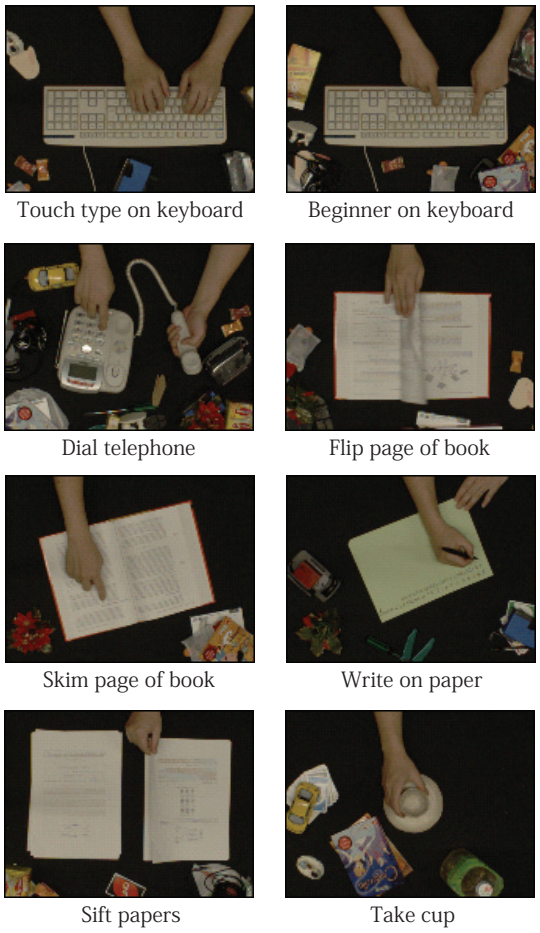

Figure 5. Examples from corpus C_{BGOB} with 8 different actions with objects and varied random background objects for each video segment.

To initialize the encoding matrix $\mathbf{H}$ for NMF near an optimal solution, we clustered the data into r clusters to compute the initial values. That is, the set of histograms $\mathbf{V}$ were clustered with k-means clustering using the top eight principle components (via PCA) of the histograms. For each column in $\mathbf{H}$, the element representing the nearest cluster was set to one and all other entries were set to random non-negative proper fractions. While in practice the final encoding matrix $\mathbf{H}$ is non-zero, zero values can be prevented using Laplacian smoothing [21].

Likewise, to initialize the EM process of the bi-modal mixture model near an optimal solution, the initial values of $p(z|d)$ are set using the results of NMF, namely, the normalized values of the encoding matrix $\mathbf{H}$ of one of the modes.

For each experiment, the number of action categories r and q are provided as prior knowledge but can also be

Table 1. Average posterior probabilities for each action category using only temporal features.

	Discovered Actions (z)							
	2	4	5	8	6	1	3	7
Touch-key	**0.97**	0.01	0.02	0.00	0.00	0.00	0.00	0.00
Begin-key	0.02	**0.82**	0.05	0.09	0.01	0.01	0.00	0.00
Dial-phone	0.02	0.00	**0.97**	0.00	0.00	0.00	0.00	0.00
Flip-page	0.00	0.00	0.00	**0.53**	0.03	0.40	0.00	0.03
Skim-page	0.00	0.00	0.02	0.00	**0.96**	0.00	0.00	0.01
Write-paper	0.01	0.48	0.27	0.09	0.03	**0.01**	0.00	0.11
Sift-paper	0.01	0.01	0.01	0.02	0.00	0.01	**0.94**	0.00
Take-cup	0.00	0.00	0.03	0.40	0.00	0.00	0.00	**0.56**

learned using a model selection criterion.

The speed of the algorithm depends on the content of the video. That being said, our algorithm on average used about 3.9 minutes to process 1 minute of video (30 fps) using a dual core 3.2 GHz CPU. On average 81.1% of the processing time was used for feature extraction, 18.8% was used for clustering and 0.1% was used for learning the bimodal model. The computational cost can be significantly reduced by parallelizing feature extraction and using online counterparts for clustering.

3.2.1 Results using only temporal features

We first performed a baseline experiment with pLSA [20] using only temporal gradient features. A bar graph of the posterior probabilities $p(z|d)$ (the probability of a latent action category z given the video segment d) computed by pLSA is shown in Figure 6. We also show the average of the posterior probabilities for each true action category in Table 1. A single row of the table is created by taking the average of the posterior probability distributions of a group of video segments from the same action. The columns of the table are ordered in such a way that maximizes the mean of the diagonal elements. We use the mean of the diagonal elements as a measure of performance, which we call the probability of correct categorization (PCC). We use this measure instead of the standard AUC measure because the AUC values for the following experiments are non-informative (i.e. they are all equal to 1).

The PCC of the motion and object corpus C_{OBJ} was 72%. The low level of performance was expected since the simple temporal gradient descriptors only capture motion and they are not invariant to scale or rotation. Nevertheless, the point is that the descriptive power of the temporal features was not sufficient to properly categorize primitive actions involving static objects. This will be true for any temporal feature descriptor based on temporal extrema.

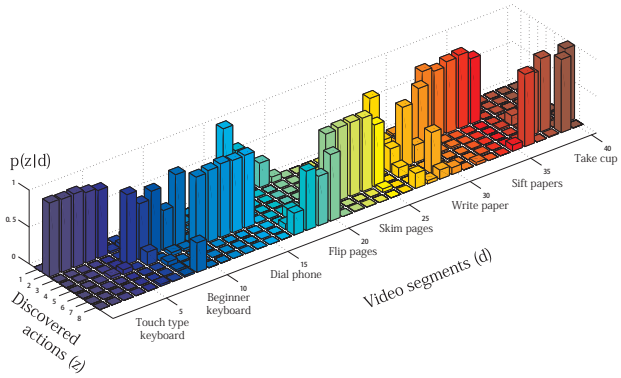

Figure 6. Baseline results: pLSA using only temporal features for corpus C_{OBJ}. The horizontal axes gives the ground truth for each video d and the discovered action category z. The vertical axis is the posterior probability $p(z|d)$.

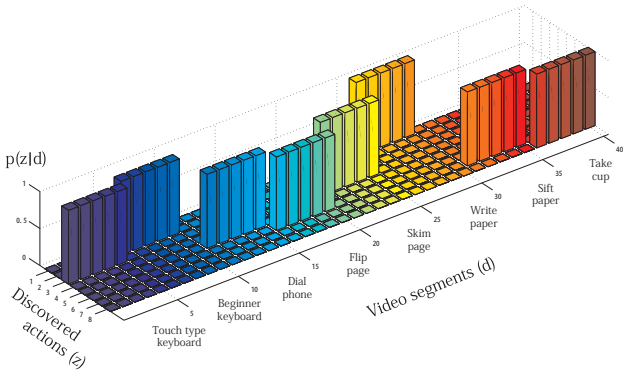

Figure 7. The resulting posterior $p(z|d)$ of corpus C_{OBJ} containing 8 different actions.

3.2.2 Learning actions using objects

Next we used our method to learn action categories from the same motion and object corpus C_{OBJ}. We observe from the bar graph of the posterior probabilities (Figure 7) that all actions contained in the video corpus have been given a high probability for the correct action category. The PCC of the corpus C_{OBJ} was 99.89%. By leveraging the visual appearance of the action and related objects, we significantly increased performance.

3.2.3 Learning actions using background appearance

The assumption of our approach is that visual context is relevant when it is constantly observed with a certain motion. We used the action and background corpus C_{BG} and tested whether our method was able to distinguish between actions

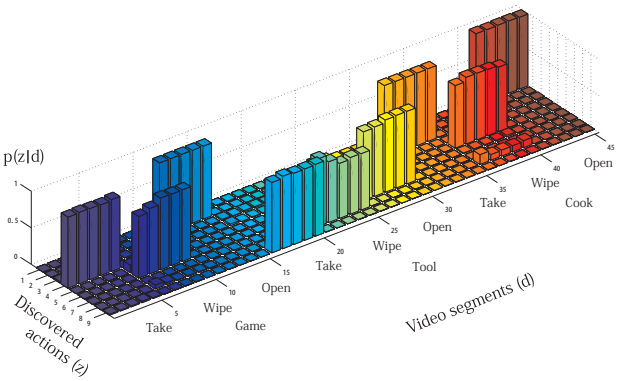

Figure 8. The resulting posterior $p(z|d)$ of corpus C_{BG} containing 3 motions and 3 backgrounds.

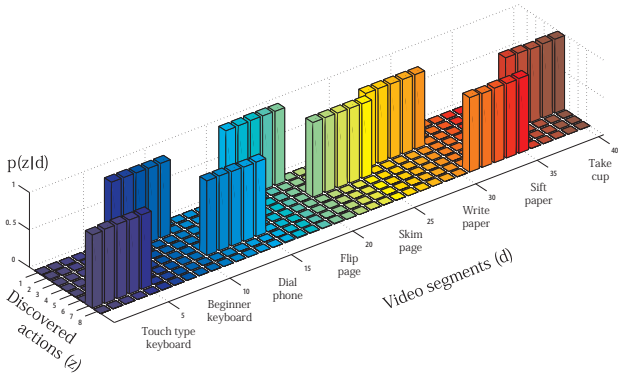

Figure 9. The resulting posterior $p(z|d)$ of corpus C_{BGOB} containing 8 actions observed over 45 various messy desktop environments.

with very similar motions, that can only be differentiated by their visual context. That is, given a video corpus containing three different motions performed in three different environments, can our method discover nine unique actions from the database?

The resulting posteriors $p(z|d)$ are given in Figure 8. Notice that each combination of visual context and motion was given a high probability for separate categories. The PCC of the corpus was 94.95%. For this corpus, the visual features induced by the combination of motion and background appearance enabled the system to differentiate the nine different action categories.

As mentioned earlier, we assume that visual context is relevant when it is constantly observed with a certain motion. We note here that our system will have problems when the number of irrelevant background types is less than the number of motion types in the database. That is, if a certain motion is frequently observed in front of the same unrelated object, our proposed method will include those visual features as part of the primitive action. In most cases, this should not be a problem given a sufficiently sized database generated over various backgrounds.

3.2.4 Learning actions using relevant visual context

In reality, primitive actions occur in various types of visual contexts and it is important to be able to leverage only the relevant visual features that should be associated with an action (Figure 1). In this experiment, we apply our proposed method to the motion with object and background corpus C_{BGOB} and show how our method can leverage relevant visual features to discover action categories, even with various cluttered backgrounds.

Our results show that our proposed approach properly discovered the action categories of the corpus C_{BGOB} by assigning high probabilities to the correct action category. The PCC was 99.38% (computed from the posterior probabilities in Figure 9). We were able to obtain good results despite the visual noise generated by the different background objects because the relevant visual features had a stronger signature (occurred more often) in the histograms; a trait of the data which was preserved by NMF. The information gained from the relevant visual features was then used by the bi-modal model to effectively discover all of the latent primitive action categories.

4 Conclusion

We have proposed a novel framework for discovering action categories by leveraging relevant visual context. To leverage visual context, we implemented a two stage clustering process via k-means clustering and non-negative matrix factorization, to generate two term-by-document matrices as the input to the bi-modal mixture model. The bi-modal mixture model used both visual features and temporal features to discover latent action categories. Our experiments showed that our approach is able to accurately categorize actions by leveraging relevant visual context to disambiguate similar motions. It was also shown that our method is robust against irrelevant visual features generated by the background while at the same time leveraging relevant visual context to accurately discover primitive action categories.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research (B) (20300061) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1402, 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *Proceedings of the International Conference on Computer Vision*, pages I:462–469, 2005.

[4] W. Buntine. Variational extensions to EM and multinomial PCA. In *Proceedings of the European Conference on Machine Learning*, pages 23–34, 2002.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] A. H. Fagg and M. A. Arbib. Modeling parietal–premotor interactions in primate control of grasping. *Neural Networks*, 11(7-8):1277–1303, 1998.

[7] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2005.

[8] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *Proceedings of the IEEE Conference on Computer Vision*, pages 1–8, 2007.

[9] R. Hamid, A. Y. Johnson, S. Batta, A. F. Bobick, C. L. Isbell, and G. Coleman. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 1031–1038, 2005.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 289–29, 1999.

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

[12] K. M. Kitani, Y. Sato, and A. Sugimoto. Recovering the basic structure of human activities from a video-based symbol string. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 9–9, 2007.

[13] I. Laptev. On space-time interest points. *International Journal on Computer Vision*, 64(2):107–123, 2005.

[14] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[15] C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[16] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, page II:1150, 1999.

[17] D. J. Moore and I. A. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *Proceedings of the National Conference on Artificial Intelligence*, pages 770–776, 2002.

[18] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 80–86, 1999.

[19] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[20] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *Proceedings of the British Machine Vision Conference*, pages III:1249–1258, 2006.

[21] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 1999.

[22] O. Okun and H. Priisalu. Fast nonnegative matrix factorization and its application for protein fold recognition. *EURASIP J. Appl. Signal Process.*, 2006(1):62–62, 2007.

[23] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[24] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the International Conference on Pattern Recognition*, pages 32–36, 2004.

[25] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[26] Y. Shi, Y. Huang, D. Minnen, A. F. Bobick, and I. A. Essa. Propagation networks for recognition of partially ordered sequential action. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 862–869, 2004.

[27] J. M. Siskind. Visual event classification via force dynamics. In *Proceedings of the National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence*, pages 149–155, 2000.

[28] T. Starner and A. Pentland. Real-time American sign language recognition from video using hidden Markov models. In *Proceedings of the International Symposium on Computer Vision*, page 265, 1995.

[29] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[30] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical Bayesian models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[31] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.