Does the Human's Representation Matter for Unsupervised Activity Recognition?

Richard G. Freedman and **Shlomo Zilberstein**
College of Information and Computer Sciences
University of Massachusetts
Amherst, MA 01003, USA
{freedman,shlomo}@cs.umass.edu

Abstract

Unsupervised activity recognition introduces the opportunity for more robust interaction experiences with machines because the human is not limited to only acting with respect to a training dataset. Many approaches currently use latent variable models that have been well studied and developed by the natural language research communities. However, these models are simply used as-is or with minor tweaks on datasets that present an analogy between sensor reading sequences and text documents. Although words have well-defined semantics so that the learned clusters can be interpreted and verified, this is not often the case for sensor readings. For example, novel data from new human activities need to be classified, which relies on the learned clusters; so how does one confirm that new activities are being correctly processed by a robot for interaction? We present several ways that motion capture information can be represented for use in these methods, and then illustrate how the representation choice has the potential to produce variations in the learned clusters.

1 Introduction

Closed-loop interactions between humans and robots have a heavy dependence on the robot's sensing capabilities. Without some form of feedback from the human and/or environment, the robot is unable to make dynamic responsive decisions with respect to the partner's newest stimuli or the consequences of its own recent actions. With the recent broad introduction of 'smart' technologies, a number of new sensors are becoming available that are affordable for everyday users, easy to wear and use ubiquitously, and connectable to other devices such as robots. Just as important as having this ability to sense is being able to effectively use the sensed information. Activity recognition is the study of how to use these low-level sensor readings to receive higher-level interpretations of the world, including users and their environments.

While supervised learning has been a well-studied approach for activity recognition (Aggarwal and Xia 2014), unsupervised learning has only gained popularity over the past decade. Its benefits for human-robot interaction are clear since humans are not restricted to precisely follow the training data, and therefore a label may not exist. So humans may perform a known action differently or execute a novel action that, though not labeled, can be clustered with proximity to a learned activity cluster. The initial work by Huỳnh, Fritz, and Schiele (2008) involved wrist sensor data and confirmed its effectiveness through comparison to a timeline of the user's activities. Since then, other works that have used unsupervised approaches for activity recognition often validate their results using either high evaluation scores with held-out testing data (Zhang and Parker 2011; Seiter et al. 2014) and/or visualizations of learned clusters that appear to justify interpretations of the dataset (Freedman, Jung, and Zilberstein 2014; Duckworth et al. 2016). However, many of these are relative or subjective methods that do not provide direct information for decision making. That is, they simply *confirm nameless labels without the inclusion of semantic explanations*. Although semantics can be added through annotation, such methods are both inefficient and defeat the purpose of unsupervised approaches.

There has been some past research investigating ways to autonomously interpret such information from unsupervised models. Because the related research we discuss from unsupervised activity recognition has frequently involved topic models on sensor data, we will particularly focus on works that study their interpretations (rather than some of the newer approaches that are being explored such as deep autoencoder neural networks (Li et al. 2014)). As a contrast to latent semantic analysis (LSA), which has been a primary tool for natural language research involving topic models (Steyvers and Griffiths 2007), Gabrilovich and Markovitch (2009) developed explicit semantic analysis (ESA) that produces descriptive features by mining relevant information. In this specific case, they identified keywords to use as features and then mined the vocabularies' Wikipedia articles for commonly associated keyword counts via term frequency-inverse document frequency (TF-IDF); this allowed text documents to be described by the keywords with greatest association (ESA) rather than by a set of word lists that were commonly in the same cluster (LSA) as the documents' words. Rather than deviate from latent semantic analysis, Kim, Rudin, and Shah (2014) created a new topic model called the Bayesian Case Model that extended the traditional latent Dirichlet allocation (Blei, Ng, and Jordan 2003) to handle inputs with discrete features - each cluster can be summarized via a 'prototype' input and its features that make it a representative of the cluster. The effectiveness of such a model is that it can use the prototypes to refine clusters during retraining, and they provided actual experiments where human users performed better in a recipe classification task using the prototype representation of clusters over the set of input lists from traditional clustering.

Unlike text documents and other naturally discrete inputs, sensor data is often continuous (omitting numerical limitations of computers, which makes the data inherently discrete) and may be interpreted in multiple ways. Each representation of the same sequence of sensor readings that may seem reasonable could yield different results. As red, green, blue-depth (RGB-D) cameras are becoming more popular (Zhang and Parker 2011; Freedman, Jung, and Zilberstein 2015; Faria et al. 2015; Duckworth et al. 2016) for both point cloud and human figure representations, we focus on the latter format of sensor data. After a brief background on topic modeling and LSA for activity recognition in Section 2, we introduce the various representations and their reasonable derivations in Section 3. Then we explore how these different representations of the same sensor data can identify different activities in Section 4. We lastly conclude with some observations, implications, and next steps in Section 5.

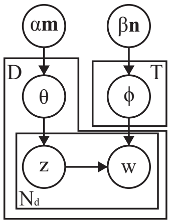
2 Background

LSA is a generative modeling approach that assumes that the observations have relevance shared by similarity in semantic meaning. However, even though each observation has its own set of meanings, it is unknown which ones relate it to the other observations. That is, there is a set of hidden definitions that describes what overall kinds of information may be observed, and identifying these definitions can explain the phenomena captured in the dataset.

From a graphical modeling perspective, these latent semantics are represented as unobserved nodes containing distributions over the set of observations. As a generative model, the 'story' explaining their presence in the dataset's creation involves sampling some distribution over distributions (such as the Dirichlet[1]) for the latent semantic distributions, and then those distributions are sampled to identify the actual observations. Learning the parameters for these distributions creates clusters of observations, and the modes of each cluster represent its semantic interpretation.

One of the most well-known applications of latent semantic analysis is topic modeling; the observations are words in a document and the latent semantic distributions are topics. Because the modes of a distribution of related words can be easily interpreted as their shared definition, topics are easily defined by the learned clusters of words. The most frequently used and tweaked topic model is Latent Dirichlet Allocation (LDA) that accounts for the fact that a document can be described topically without needing to know the order of the words (Blei, Ng, and Jordan 2003). Its graphical model, generative process, and mathematical representation are presented in Figure 1.

```
Algorithm 1: GenerateLDA($T$, $D$, $N$, $\alpha\vec{m}$, $\beta\vec{n}$)
forall the topic $t \in \{1, \ldots, T\}$ do
    draw $\phi_t \sim$ Dirichlet $(\beta\vec{n})$
forall the document $d \in \{1, \ldots, D\}$ do
    Draw $\theta_d \sim$ Dirichlet $(\alpha\vec{m})$
    forall the index $i \in \{1, \ldots, N_d\}$ do
        Draw topic $z_{d,i} \sim \theta_d$
        Draw word token $w_{d,i} \sim \phi_{z_{d,i}}$
```



$$\mathrm{P}\left(\vec{z}, \vec{\theta}, \vec{\phi} \,|\vec{w}, \alpha\vec{m}, \beta\vec{n}\right) \propto \mathrm{P}\left(\vec{w} \,\middle|\vec{z}, \vec{\phi}\right) \cdot \mathrm{P}\left(\vec{z} \,\middle|\vec{\theta}\right) \cdot \mathrm{P}\left(\vec{\theta} \,|\alpha\vec{m}\right) \cdot \mathrm{P}\left(\vec{\phi} \,|\beta\vec{n}\right)$$

$$\text{where } \mathrm{P}\left(w_{e,j} \in \vec{w} = v \,\middle|\vec{z}, \vec{\phi}\right) \approx$$

$$\left(\sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbf{1}\left(w_{d,i} = v \wedge z_{d,i} = z_{e,j} \wedge (e,j) \neq (d,i)\right) + \alpha m_v\right) /$$

$$\left(\sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbf{1}\left(z_{d,i} = z_{e,j}\right) - 1 + \alpha\right) \text{ and}$$

$$\mathrm{P}\left(z_{d,j} \in \vec{z}_d = t \,\middle|\vec{\theta}\right) \approx \left(\sum_{i=1}^{N_d} \mathbf{1}\left(z_{d,i} = t \wedge j \neq i\right) + \beta n_t\right) / \left(N_d - 1 + \beta\right)$$

Figure 1: LDA's generative process, graphical model, and mathematical representation.

Activity recognition research that uses vanilla LDA has to assume that in a similar manner, describing what generally happens during a sequence of sensor readings does not rely on the ordering. Then the clusters of sensor readings can be interpreted by the modes, if interpretation is desired at all. However, temporal consistency has been acknowledged and considered with LDA and similar topic models. Zhang and Parker (2011) applied a sliding window over sequences of RGB-D point clouds and compressed the large amount of observed data into smaller vectors, and Freedman, Jung, and Zilberstein (2015) used a topic model that includes LDA within a hidden Markov model called the composite model (Griffiths et al. 2004). The former approach embedded the *temporal information within the observation* while the latter approach applied a *temporal aspect to the model.*

3 Representations of Human Posture Data

Human posture data can be read today using a variety of sensors, including traditional motion capture (Hodgins 2002), RGB-D camera point cloud analysis (Shotton et al. 2011), and IMU sensors attached to the body (Jung et al. 2015). Using the links and/or joints of the body, the sensor data is able to render three-dimensional stick figures of the observed agent - this is often done through computation of the homogeneous matrix transformations between adjacent links of the body.

Depending on the number of links that the sensor can identify, each reading can be very high-dimensional. Therefore, due to both this and the continuous space of transformations, the data must be compacted to be useful with a reasonable amount of data collection. The opportunity to compress the data not only lies within each frame, but also between frames. We consider various cases that will compress the human posture data differently, but also present the merits for why it could be the best choice for representation of

[1] The Dirichlet is a random probability mass function that captures the variance of distributions identified from observed events that actually occur based on some true distribution (Frigyik, Kapila, and Gupta 2010). For example, flipping a fair coin should produce heads half the time when observed over many trials, but getting heads forty-nine percent of the time would also not be too surprising. However, getting heads ten percent of the time is unlikely.

observations in latent semantic analysis.

Independence between Frames

As each sensor reading corresponds to a single frame of animation, it is the most indivisible unit of the sequence. Although the text analogy is not exact in this case because words can be broken down semantically into phonemes and further down without semantics into orthographic symbols (letters, numbers, etc.), this is the simplest input and carries a reasonable amount of semantic information by itself. Just as the adage “a picture can be worth 1000 words” implies, one can describe a single frame’s posture with various feature assignments; this was slightly explored in an attempt to use ESA to describe the modes of clusters learned via LSA (Freedman and Zilberstein 2016). One advantage of running an activity recognition model frame-by-frame is for *on-line artificial intelligence used during interaction* - every sensor reading may, but not necessarily has to, be used right away for decision making purposes.

Joint Angles Joint Angles are the rawest format for the posture representation. Often provided as a triplet of Euler angles per joint, but sometimes in quaternions to avoid gimbal lock (removing ambiguity when visualizing the angle), these values derive the matrix rotation transformations that render the stick figure. Translations come from the link lengths, which vary per individual, but the rotations are generally independent of the individual. This makes them ideal for *generalizing the learned activity models to other people* who may interact with the robot, and it allows training data to come from multiple subjects. It is also the reason that the relative distance or position of points in Cartesian space is less ideal for representing the stick figure instead. In particular, the additional information of the actual points in space is not very useful for representing other individuals while the joint angles that can be generalized to others require more steps to extrapolate.

Derivatives Just as the rawest format of a posture is the joint angles, the rawest format between consecutive postures is the derivative. Motion is a continuous process, which has been used to find keyframes in motion capture videos and present them in comic book form with ‘whoosh’ lines to summarize the motions (Choi et al. 2012). This allows the temporal aspect to be embedded within the representation to a very small extent, but still keep the dimensionality much smaller than through the use of a sliding window. By computing approximations of the derivative as the difference between joint angles in consecutive frames, each derivative is still easy to obtain and then use right away. However, without a frame of reference such as a starting joint angle configuration, these derivatives can be very vague and lead to ambiguity during clustering. For example, raising the arm above the head from in front will be identical in this representation to raising the arm to the front from the side.

Features As briefly mentioned above, it is possible to extract features mathematically from a single human posture. In particular, the posture was broken down into labeled features such as whether each arm was bent, the back was straight, a leg was raised, etc.. If such information is already available, then it may be worth clustering feature vectors directly so that one may inspect the shared features between the modes of a cluster. Furthermore, because many of the features have finite possibilities (e.g. the arm is above the head, at the side, or in front of the body), it is possible to take advantage of the Bayesian Case Model that broke its inputs/observations down in a similar way (Kim, Rudin, and Shah 2014). A prototype would be a posture with specific parts of the body in fixed relative positions (joint angle) or simple motions (derivative) while the rest the body remains flexible.

This is different from unsupervised feature learning (Li et al. 2014), which employs unsupervised learning techniques on a dataset to get clusters as lower-dimensional features. These clusters are used to aid in supervised learning tasks; thus unlabeled features are obtained for classification with labeled classes. Instead, we propose extracting lower-dimensional labeled features from data to use for classification without labeled classes.

Parametric In contrast to the above representations that view each frame as a single observable input, it is possible to break the information down into smaller pieces. For example, each joint’s angles or derivatives may be viewed as an individual triplet (versus concatenating all the joints’ together) and each feature may also be inspected independently. Although such perspectives cannot be handled by traditional topic models that use single word inputs, we may adjust the generative models with new ‘stories’ that better explain the generation of human postures for various activities. One such model that may be of use for this is Parameterized LDA (Freedman, Jung, and Zilberstein 2015), which runs LDA with a single latent topic across multiple vocabularies simultaneously. In this case, each vocabulary may seem identical such as the interval triplet $[-\pi, \pi]^3$ for a single joint angle, but it would allow the distributions to be more informative about how each joint is involved in an activity’s description. A uniform distribution over this space would imply that the joint has no significance because its high entropy is not discerning, but a multi-modal peak would imply that some angles for this joint are more commonly associated with the clustered activity.

Multimodal The increase in amount and types of sensors has also led to a boom in multimodal learning. This allows synergy because one sensor may identify things that can complement the information missing from another sensor. Besides using additional sensors to provide more information than the human’s posture, we may also consider *combining multiple representations of the same data*. Because they all have different interpretations and a strong chance of recognizing different activities (see Section 4), we should consider taking advantage of multiple perspectives at once to get the ‘complete picture’ at each frame. However, it is also important to consider the dimensionality of the data because the space of inputs will grow drastically as more representations are used simultaneously.

Joint Relationships between Frames

With the concern over whether a single frame truly represents any semantic context of an activity, it is possible to consider a collection of frames at once. This produced the temporal components described at the end of Section 2 to go with the spatial representations. Although this does contain richer context, it comes with the prices of more computation time to acquire the data (a potential bottleneck in online interaction) as well as the curse of dimensionality. The number of parameters is linear with respect to the number of joined frames, but a single frame already has many variables - the ones described above range from approximately 30 to 45 without the multimodal representation. To cope with this, the training data is compressed a priori to form a codebook, which is a discriminative unsupervised clustering of the observations. This codebook maps each observation to its cluster ID as the input for LSA, but then the challenge of how many clusters to include in the codebook arises. Due to space limitations, we simply note that the representations for multiple frames are identical to the ones for single frames except that we either concatenate or apply some averaging formula over a sliding window of observations.

4 Impacts of Representation Choice

Although all the representations discussed in Section 3 come from the same stream of sensor readings, they portray different aspects of the observations. However, should they still be able to recognize the same activities? In this section, we will illustrate impacts with respect to both LSA for recognition and available information for later decision making.

Theoretically, there is *no guarantee of learning the same activities over multiple clustering runs with the same representation choice* due to the label switching problem, which states that the clusters learned via mixture models such as LSA are equally likely to be any permutation of the optimal latent variable assignments (Stephens (2000) provides a good explanation of the problem and some attempts to address it). However, even from the algorithmic perspective where a random seed can replicate results if held constant, will there still be clusters for the same activities when running LSA on different representations of the same data with a constant random seed?

To simply illustrate how LSA clusters semantically, we will consider two separate text phrases: “The dog jumped over the fence” and “Running to buy a running refrigerator”. If they are in the same dataset, then the words will be viewed as ten different symbols and is not relevant to our discussion. However, if they are each a separate dataset, just as we abstract each representation of the sensor reading, then the phrases may be encoded using the same five symbols: 1 2 3 4 1 5. This makes the two phrases *appear identical symbolically*, and using the same random seed when running them in a MCMC algorithm (such as Gibbs Sampling, which iterates over each symbol in sequential order (Griffiths 2002)) for clustering under LSA will result in clustering them the same way. The only difference between these clusters will be the list of modes, which are decoded into the actual words.

Although the example is an oversimplification because this phenomena is less likely with much larger datasets of text, should we accept that this will also not happen with different representations of the same sensor data from the single observed activity sequence? As a brief example that different representations of the same sensor data would not align symbolically, suppose a person is lifting and putting down boxes; then the observations will have postures where the arms are rising and then lowering. This sequence has symmetry over the joint angles with a symbolic representation in the form of 1 2 3 ... $(k-1)$ k $(k-1)$ ... 3 2 1, but the derivatives will lack this symmetry because of the change in direction of movement to yield symbolic representation 1 2 3 ... $(k-1)$ k $(k+1)$ ... $(2k-1)$. There may be some consecutive duplicates in the derivative's representation if the acceleration happens to ever be zero, but this is still drastically different from the joint angles' symbolic representation when processed sequentially for clustering by MCMC. Likewise, if sliding windows are applied and a codebook is necessary, then there may be yet another unique symbolic representation of the same sequence of sensor readings. This will, even with the same random seed, algorithmically produce different clusters that may end up with different interpretations. Thus we must take this into consideration because, though an activity will be described differently from each representation's perspective, *are we guaranteed that these interpreted clusters will actually describe the same activity*?

In addition to this potential difference, the representation choice will impact the information available for responsive decision making. Joint angles alone, for example, yield posture without any insight into the human's upcoming position. This may complicate prediction for collision avoidance and supports the use of a derivative representation. However, abstracting to just the derivative representation removes the local posture. Because all this information is initially available from the sensor data, we propose abstracting the correct representation(s) for the decision making task even if they are different from what is used for recognition.

5 Discussion

In human-robot interaction, autonomous response requires some degree of recognizing what partners in the interaction are doing. We have discussed how even for a single sensor, there are multiple representation choices that can affect either the observation format, the model, or the algorithm. Besides those effects, each representation can have very different underlying forms that may lead to the risk of altering the performance of LSA in unsupervised activity recognition.

Does this mean a single representation is correct and the others are wrong? Do we need to try them all and then use the one that is best with each learning session? Are there evaluation metrics we can develop to make such decisions? If we can identify when each representation is best to use, then perhaps we need to consider metareasoning during interactive decision making to decide when to use each one. We are exploring these issues and starting an empirical evaluation of each representation's learned clusters, comparing them with respect to current LSA evaluation metrics.

Acknowledgments

The authors would like to thank the reviewers for their thorough feedback and insights. This work was supported by NSF grant number IIS 1405550.

References

Aggarwal, J. K., and Xia, L. 2014. Human activity recognition from 3D data: A review. *Pattern Recognition Letters* 48:70–80.

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

Choi, M. G.; Yang, K.-Y.; Igarashi, T.; Mitani, J.; and Lee, J. 2012. Retrieval and visualization of human motion data via stick figures. *Computer Graphics Forum* 31(7):2057–2065.

Duckworth, P.; Gatsoulis, Y.; Jovan, F.; Hawes, N.; Hogg, D. C.; and Cohn, A. G. 2016. Unsupervised learning of qualitative motion behaviours by a mobile robot. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, 1043–1051. Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems.

Faria, D. R.; Vieira, M.; Premebida, C.; and Nunes, U. 2015. Probabilistic human daily activity recognition towards robot-assisted living. In *Proceedings of the Twenty-Fourth IEEE International Symposium on Robot and Human Interactive Communication*, 582–587.

Freedman, R. G., and Zilberstein, S. 2016. Using metadata to automate interpretations of unsupervised learning-derived clusters. In *First IJCAI Workshop on Human is More Than a Labeler (Beyond-Labeler)*, 1–7.

Freedman, R. G.; Jung, H.-T.; and Zilberstein, S. 2014. Plan and activity recognition from a topic modeling perspective. In *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling*, 360–364.

Freedman, R. G.; Jung, H.-T.; and Zilberstein, S. 2015. Temporal and object relations in unsupervised plan and activity recognition. In *Proceedings of AAAI 2015 Fall Symposium on AI for Human-Robot Interaction*, 51–59.

Frigyik, B. A.; Kapila, A.; and Gupta, M. R. 2010. Introduction to the dirichlet distribution and related processes. Technical Report UWEETR-2010-0006, University of Washington, Seattle, WA, USA.

Gabrilovich, E., and Markovitch, S. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* 34(1):443–498.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems*, 537–544.

Griffiths, T. L. 2002. Gibbs sampling in the generative model of latent Dirichlet allocation. Technical report, Stanford University.

Hodgins, J. K. 2002. Carnegie Mellon Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu. [Online, last viewed August 2017].

Huỳnh, T.; Fritz, M.; and Schiele, B. 2008. Discovery of activity patterns using topic models. In *Proceedings of the Tenth International Conference on Ubiquitous Computing*, 10–19.

Jung, H.-T.; Freedman, R. G.; Takahashi, T.; Wong, J. M.; Zilberstein, S.; Grupen, R. A.; and Choe, Y.-K. 2015. Adaptive therapy strategies: Efficacy and learning framework. In *Proceedings of the IEEE/RAS-EMBS International Conference on Rehabilitation Robotics*, 950–955.

Kim, B.; Rudin, C.; and Shah, J. A. 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems Twenty-Seven*. Curran Associates, Inc. 1952–1960.

Li, Y.; Shi, D.; Ding, B.; and Liu, D. 2014. Unsupervised feature learning for human activity recognition using smartphone sensors. In Prasath, R.; O'Reilly, P.; and Kathirvalavakumar, T., eds., *Proceedings of the Second International Conference on Mining Intelligence and Knowledge Exploration*, 99–107. Cork, Ireland: Springer International Publishing.

Seiter, J.; Derungs, A.; Schuster-Amft, C.; Amft, O.; and Tröster, G. 2014. Activity routine discovery in stroke rehabilitation patients without data annotation. In *Proceedings of the Eighth International Conference on Pervasive Computing Technologies for Healthcare*, 270–273.

Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; and Blake, A. 2011. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 1297–1304.

Stephens, M. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62(4):795–809.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. In Landauer, T.; McNamara, S. D.; and Kintsch, W., eds., *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Zhang, H., and Parker, L. E. 2011. 4-dimensional local spatio-temporal features for human activity recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2044–2049.