A GENERALIZED DATA-DRIVEN HAMILTONIAN MONTE CARLO FOR HIERARCHICAL ACTIVITY SEARCH

Ricky J. Sethi[⋆] *Hyunjoon Jo*[⋆] *Amit K. Roy-Chowdhury*[†]

[⋆] USC Information Sciences Institute
[†] University of California, Riverside

ABSTRACT

Motion and image analysis are both important for robust solutions to video search of activities; the physics-based, data-driven Hamiltonian Monte Carlo (HMC), a Markov chain Monte Carlo variant that is efficient in searching large dimensional spaces, simultaneously examines the combined motion and image space. In this paper, we generalize the data-driven HMC to no longer depend upon ad hoc Guide Hamiltonians and to no longer require physics-based features from tracks as pre-requisites. Our generalization thus allows it to be used with or without a tracker, overcoming a significant limitation of the physics-based approach, as well as being extensible to utilizing any pre-existing image- or motion-based method. We demonstrate the generalizability of our framework by considering situations when tracking is available and when it is not available. When tracking is available, we utilize Histogram of Oriented Gradients, shapes of trajectories, and Hamiltonian Energy Signatures; when tracking is not available, we use Space-time Interest Points and GIST features. In addition, we show our generalized framework performs better than the physics-based, data-driven HMC, as well as state-of-the-art, by demonstrating the efficacy of our system on real-life video sequences using the well-known Weizmann and YouTube Action datasets.

Index Terms— Stochastic Integration, Hamiltonian Monte Carlo, Data-Driven

1. INTRODUCTION

The dynamic nature of video makes activity search and recognition from video databases a very difficult problem [1, 2, 3, 4, 5]. It requires an analysis of both the motion and the image features of the system being studied in video. These often disparate features need to be combined in order to realize optimal recognition. Their integration, however, leads to an even greater problem as the final search space, which is composed of the combined motion and image spaces, is usually enormous and extremely complex.

There have been some approaches [6, 7, 2, 8] that have considered both the space of motion features and the space of image features but these techniques depend on supervised methods and require significant training. However, statistical sampling techniques like Markov chain Monte Carlo (MCMC) have proven especially effective at analyzing complex spaces, such as the combined motion and image spaces. Hamiltonian/Hybrid Monte Carlo (HMC) is an MCMC variant that uses gradient information to make traditional MCMC more efficient by leveraging the advantages of Hamiltonian dynamics to investigate how the system evolves in parameter space. This gives the HMC higher acceptance rates, less correlated and faster converging chains, and suppression of the random walks in traditional MCMCs [9, 10]. Rather than utilizing the blind proposals in a traditional HMC, the data-driven HMC [11] uses trajectory information gleaned via a tracker to create physics-based features which are then used as informed proposals for the HMC framework. The physics-based, data-driven HMC explores both motion and image spaces by creating data-driven proposals in motion space and then confirming them in the image space, thus providing a systematic approach that allows understanding of the effects of each of the feature spaces separately. However, its use of so-called Guide Hamiltonians and its reliance on physics-based features from trajectories derived via a tracker limits its utility and applicability greatly.

In this paper, we generalize the data-driven HMC to remove the Guide Hamiltonians restriction as well as to remove the reliance on physics-based trajectory information. Our generalization thus allows it to be used with or without a tracker, overcoming a significant limitation of the physics-based approach, as well as being extensible to utilizing any pre-existing image- or motion-based method instead of relying upon physics-based features only. We demonstrate the generalizability of our framework by considering situations when tracking is available and when it is not available. When tracking is available, we utilize Histogram of Oriented Gradients [12], shapes of trajectories [13], and Hamiltonian Energy Signatures [14, 15, 16] (the last two as utilized in the original data-driven HMC implementation [11]); when tracking is not available, we use Space-time Interest Points [17] and GIST [18, 4] features. In addition, we show our generalized framework performs better than the physics-based, data-driven HMC, as well as state-of-the-art, by demonstrating the efficacy of our system on real-life video sequences us-

ing the well-known Weizmann and YouTube Action datasets. Finally, our generalized, data-driven HMC can utilize either image or motion space proposals, as opposed to being limited to only using motion-based proposals as in the physics-based data-driven HMC. We are thus able to bring the full power of the data-driven HMC framework to video analysis.

Algorithm 1 Generalized Data-Driven HMC (DDHMC)

Here, v_{query} is the query video and v_i is the i^{th} input video clip from the database. $D_{Motion}(v_i, v_{query})$ is the distance between the query video and the i^{th} input video clip from the database based on Motion information and $D_{Image}(v_i, v_{query})$ is the distance based on Image information. $P_{Image}(D_q)$ finds the video clip with the distance score closest to the distance score, D_q, in the Image space. $K(D(v_i, v_{query}))$ is the density estimator. Initialize the chain with Δt, $q_o \sim e^{-K(D_{Image}(v_o, v_{query}))}$, and $v_o = P_{Image}(D_{q_o})$.

for $i = 1$ to *nsamples* **do**
  // **Step 1**. Data-Driven step
  flag = true
  while (flag) **do**
    draw $v_i^{'} \sim e^{-K(D_{Motion}(v_i, v_{query}))}$
    draw $\alpha \sim \mathcal{U}[0,1]$
    if $\alpha > min\left(1, \frac{D_{Motion}\left(v_i^{'}, v_{query}\right)}{D_{Motion}(v_{i-1}, v_{query})}\right)$ **then**
      flag = false
    end if
  end while
  $p = D_{Motion}(v_i^{'}, v_{query})$
  $(q^o, p^o) = (q_{i-1}, p)$
  // **Step 2**. Dynamic Transition Step (LeapFrog)
  for $j = 1$ to L **do**
    $p^{j-\frac{1}{2}} = p^{j-1} - \frac{\Delta t}{2} \bullet \nabla U(q^{j-1})$
    $q^j = q^{j-1} - \Delta t \bullet p^{j-\frac{1}{2}}$
    $p^j = p^{j-\frac{1}{2}} - \frac{\Delta t}{2} \bullet \nabla U(q^j)$
  end for
  $v^L = P_{Image}(D_{q^L})$
  $q^L = D_{Image}(v^L, v_{query})$
  $(q^{'}, p^{'}) = (q^L, p^L)$
  // **Step 3**. Final Metropolis-Hastings Step
  draw $\alpha \sim \mathcal{U}[0,1]$
  $\delta H = H(q^{'}, p^{'}) - H(q^o, p^o)$
  if $\alpha < min\left(1, e^{-\delta H}\right)$ **then**
    $(q_i, p_i) = (q^{'}, p^{'})$
  else
    $(q_i, p_i) = (q_{i-1}, p_{i-1})$
  end if
end for
return $\{q_i, p_i\}_{i=0}^{nsamples}$

2. IMAGE AND MOTION SPACE INPUTS

The inputs to the generalized data-driven HMC are the Image space probability density function (pdf) ($\pi(f)$) and the Motion space pdf ($\pi(\tau)$). We compute these using a density estimator (Section 2.3) on the Image and Motion space similarities calculated between the query clip and every clip in the test database. Since we only need similarities, our construction provides flexibility on the particular methods employed to calculate either Image or Motion features; usually, one space can give a broad idea but detailed analysis has to be in the other space. As such, we can use well-established methods in computer vision to calculate image or motion features since such analysis is a well-known area in activity recognition [1], instead of being constrained to only using physics-based trajectory features as in the physics-based data-driven HMC. We demonstrate the versatility and extensibility of our framework by employing two different approaches for both the Motion space and Image space calculation in these experiments, as detailed below.

2.1 The Motion Space

We consider two situations here: one, where tracking is available and two, where tracking is not available. If tracking is feasible on a dataset (e.g., on the Weizmann dataset), we use the object detector from [19] to ensure the video is segmented into objects and their motion is given. Then, from the physical motion and location information of objects over time, we can utilize a plethora of motion features based on trajectories. In particular, we compute similarity scores between the computed trajectories as there are many ways this can be done (e.g., [13]). If tracking is not feasible on a dataset (e.g., on the YouTube Action dataset), we follow the example of [17] and use Space-time Interest Points, which does not rely on motion segmentation or other pre-processing steps, in combination with the Nearest Neighbour classifier, as detailed in [20], to finally yield the similarities.

2.2 The Image Space

Similarly to the Motion space formulation, we use well-established methods in computer vision to calculate image-based features for our representation. Our construction provides flexibility since new approaches in low-level feature extraction can be employed easily within our framework. In general, we can follow the example of other data-driven MCMC approaches like [21, 22, 23].

For the present work, we again consider the cases when tracking is available (Weizmann) or not (YouTube Action). If it is available, we use the shape of the feature points in the video, which can be the shape of a trajectory or the shape of our object, in the analysis of the Weizmann dataset. Distances can be computed between shapes, leading to a similarity matrix [13]. We can also use Histogram of Oriented Gradients [12], trajectory-based descriptors, colour/texture, etc., since the integration is directly on the similarity scores. For analysis of the YouTube Action dataset, we utilize the GIST [18] to create global scene features for each frame of a video and then average each of the features for that video as a whole, over all the frames for that video. We then compute the distance between the averaged global features for each video with the

query and use that to determine the similarity by employing a Bag of Words [24] model with a k-means classifier.

2.3 Proposal Distribution Formulation

We then cast the similarity scores from our Motion and Image spaces as a Gibbs proposal distribution since any distribution that is nowhere zero can be put in a canonical (Gibbs) distribution form [10, 25]. In order to estimate the distribution for the similarity scores, we follow [11] and use standard Kernel Density Estimation (KDE) [26]:

$$K(D) = \frac{1}{nh}\sum_{i=1}^{n} K_{eff}\left(\frac{D-d_i}{h}\right)$$
$$\text{with } K_{eff}(D) = (2\pi)^{-\frac{1}{2}} e^{-\frac{D^2}{2}} \text{ and } d_i \in (D-h, D+h] \quad (1)$$

where D is the distance measure between two tracks and h is the bandwidth, which is set using k-Nearest Neighbour, as described in [26].

3. OVERVIEW OF THE GENERALIZED DDHMC

Starting with a query video clip and a database of test clips, we first analzye each test clip against the query clip using both image and motion analysis, resulting in similarity distributions for both the image and the motion space, as described in Section 2. We then convert these similarities to probability distribution functions as shown in Section 2.3.

Once we have the distributions for both image and motion evaluation, we are ready to provide the input to the full generalized data-driven HMC algorithm, as shown in Algorithm 1. We create the initial position variable, q_o, by sampling from the Image space, which we subsequently confirm in the Motion space (we could just as well get the initial sample from the Motion space, in which case we would then confirm in the Image space). This q_o is exactly what we use to find the initial video clip, v_o, and we send both of them into Step 1, below. The q_i thus become the Image-based proposals we try to confirm using the generalized data-driven HMC framework.

Like the physics-based, data-driven HMC, our integration affords a hierarchical classification scheme in which the data-driven proposal does an initial, gross classification not possible in the traditional HMC. However, unlike the physics-based, data-driven HMC, we do not need to create ad hoc "Acceptance" or "Proposal/Guide" Hamiltonians and do not require physics-based features from tracks as pre-requisites.

4. EXPERIMENTS

We conduct experiments on the Weizmann, YouTube Action, and Berkeley Segmentation datasets to demonstrate how the integration afforded by the generalized data-driven HMC outperforms the physics-based data-driven HMC, the Traditional HMC, and the state of the art. Also, like the physics-based approach, the generalized data-driven HMC helps reduce the search space using the data-driven portion, as well as the hierarchical scheme for recognition. We also utilize these datasets to show the flexibility of the generalized data-driven HMC framework to accomodate any method to compute the Image and Motion features by considering the case when tracking is available (Weizmann), as well as when it is not (YouTube Action and Berkeley Segmentation), as detailed in Section 2.

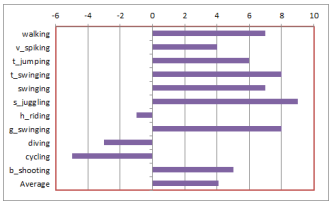

Fig. 1. Differences for each YouTube dataset activity between [27] and the Generalized Data-Driven HMC showing average improvement of 4.1%.

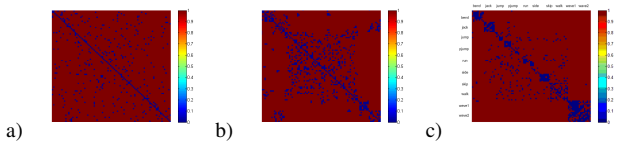

Fig. 2. Similarity matrices using the Weizmann dataset for a) Traditional HMC Integration [10], b) Physics-Based Data-Driven HMC Integration [11], and c) Generalized Data-Driven HMC Integration.

4.1 Integration on the Weizmann Dataset

The Weizmann dataset [28] consists of a database of 90 low-resolution (180 x 144, deinterlaced 50 fps) video sequences showing nine different people, each performing 10 natural actions. In Figure 2, we compare similarity matrices using the Weizmann dataset for a) the Traditional HMC, b) the Physics-Based Data-Driven HMC, c) and the Generalized Data-Driven HMC Integration. The rows and columns represent 10 activities by people and are organized according to activity. The plots show the clarification of matches using the different methods: in (a), the Traditional HMC tends to have no discernible pattern of matching; in (b), the Physics-Based Data-Driven HMC tends to do a little finer granularity of classification; but (c) the Generalized Data-Driven HMC shows finest granularity and distinction of matches and classification. As in the approach of [29], the comparison to Weizmann is not intended to show absolute improvement, as that has already been shown to be maximized by other methods, but to show relative improvement over other methods on a common dataset; i.e., to show the improvement in previously poorly performing methods when fused via our Generalized Data-Driven HMC. In addition, the Generalized Data-Driven HMC provided a 23% improvement in accuracy over the Physics-Based Data-Driven HMC.

4.2 Integration on the YouTube Action Dataset

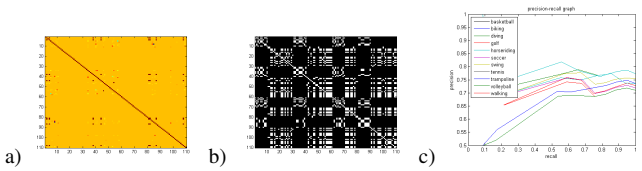

Fig. 3. (a) Input Motion similarity distributions; (b) Input Image and (c) Precision/Recall curves of the Integration via the Generalized Data-Driven HMC: averages of query examples' Precision/Recall curves (for each class).

We also look at the YouTube Action dataset [27]. The YouTube Action dataset contains 11 action categories and is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, etc. Finally, with the YouTube Action Dataset, similar to the performance on Weizmann, we see the Generalized Data-Driven HMC is able to identify the main activity without confusing it with any other activity and that the image method narrows in on the general class while the motion method achieves the final categorization of the hierarchical search. We show precision recall curves, and the input similarity matrices, for the YouTube Action analysis in Figure 3, in which we used GIST plus Bag of Words for the Image method and STIP plus SVM for the Motion analysis, as detailed in Section 2. The precision recall curves are the averages of query examples' precision recall curves (for each class). We computed these for each of the 11 classes with 10 examples from each single class. For each of the 10 examples, we computed matching with each of the 11 classes and then averaged these precision recall values.

In these experiments, we have just used the basic STIP features since specialized feature sets is not the focus of this work. Also, these results should not be compared to absolute recognition scores, but rather as the gain over the image-based approach and the pruning of the search space in our hierarchical approach. In addition, we also compared it to the original implementation on the YouTube dataset [27] and, based upon our implementation of their methodology using approximately 400 cuboids and between 6,000-11,000 static features from each video without pruning, achieved an average improvement of 4.1% as shown in Figure 1, which is comparable to the improvement in [29]. The Generalized Data-Driven HMC outperformed in most categories except h_riding, diving, and cycling where the contribution from the input motion similarity was especially poor. And, as previously mentioned, the Physics-Based Data-Driven HMC [11] could not even be applied to this dataset.

4.3 Comparison to State-of-the-Art

We also compared our method to state of the art sampling methods like MCMC [30], Reversible Jump MCMC (RJMCMC) [31], and HMC [10]; in order to compare to previously published methods and show the wide range of the Generalized Data-Driven HMC over the Phyiscs-Based Data-Driven HMC, we use the well-known problem of segmentation of colour images and compare to [31, 32, 30, 10] in Table 1. Following the example of [31], we also utilized the Berkeley Segmentation Dataset [33]. The Generalized Data-Driven HMC outperforms them and, once again, the Physics-Based Data-Driven HMC [11] could not even be used here.

Method	F-Measure
Ours	0.61
RJMCMC [31]	0.57
JSEG [32]	0.56
HMC [10]	0.41
MCMC [30]	0.34

Table 1. Comparison of Generalized Data-Driven HMC ("Ours") to previously published sampling methods.

These dataset results show the potential for wide applicability of the Generalized Data-Driven HMC framework to many different modalities, in addition to significantly reducing the search space in video database search problems. Since activity search in video is becoming a very important problem, we expect the Generalized Data-Driven HMC to be an important contribution in this direction.

5. CONCLUSION

We presented a Generalized Data-Driven HMC (DDHMC) that extends the Physics-Based Data-Driven HMC to no longer require ad hoc Guide Hamiltonians or physics-based track information. This generalization beyond its reliance on only physics-based, heuristic features allows our Generalized DDHMC to be used with or without a tracker, overcoming a significant limitation of the physics-based approach, as well as being extensible to utilizing any pre-existing image- or motion-based method. We demonstrated the generalizability of our framework by considering situations when tracking is available and when it is not available. Our generalized framework allows the HMC to be applied to a vast variety of situations which were not possible with the Traditional HMC or the Physics-Based Data-Driven HMC and extends the Data-Driven HMC to any general stochastic problem, thus giving the ability to use the Data-Driven HMC as a fusion methodology for general stochastic problems.

6. REFERENCES

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *CSVT*, 2008.

[2] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, 2012.

[3] R J Sethi, N Nayak, and A K Roy-Chowdhury, *Motion Pattern Analysis for Event and Behavior Recognition*, Springer, 2011.

[4] Nitesh Shroff, Pavan Turaga, and Rama Chellappa, "Moving vistas: Exploiting motion for describing scenes," CVPR, 2010.

[5] Bi Song, Ricky J Sethi, and Amit K Roy-Chowdhury, "Robust Wide Area Tracking in Single and Multiple Views," in *Visual Analysis of Humans*, T B Moeslund, L Sigal, V Krüger, and A Hilton, Eds., pp. 1–18. Springer-Verlag, 2011.

[6] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *CVPR*, 2009.

[7] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *PAMI*, vol. 31, 2009.

[8] R.J. Sethi, A.K. Roy-Chowdhury, and S. Ali, "Activity recognition by integrating the physics of motion with a neuromorphic model of perception," *WMVC*, 2009.

[9] M. Alfaki, "Improving efficiency in parameter estimation using the hamiltonian monte carlo algorithm," M.S. thesis, University of Bergen, 2008.

[10] R.M. Neal, "Probabilistic inference using markov chain monte carlo methods," Tech. Rep., University of Toronto, 1993.

[11] R.J. Sethi and A.K. Roy-Chowdhury, "A neurobiologically motivated stochastic method for analysis of human activities in video," *ICPR*, 2010.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005.

[13] A. Veeraraghavan, A.K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human motion analysis," *PAMI*, 2005.

[14] R.J. Sethi, A.K. Roy-Chowdhury, and A. Veeraraghavan, "Gait Recognition Using Motion Physics in a Neuromorphic Computing Framework," in *Multibiometrics for Human Identification*. CUP, 2010.

[15] Ricky J Sethi and Amit K Roy-Chowdhury, "Modeling and Recognition of Complex Multi-Person Interactions in Video," in *ACM MM MPVA*, 2010, pp. 0–3.

[16] Ricky J Sethi and Amit K Roy-Chowdhury, "Physics-based Activity Modelling in Phase Space," in *ICVGIP*, 2010.

[17] I. Laptev and P. Perez, "Retrieving actions in movies," *ICCV*, 2007.

[18] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *IJCV*, vol. 42, pp. 145–175, 2001.

[19] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," *CVPR*, 2008.

[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," *ICPR*, 2004.

[21] Z. Tu, S. Zhu, and H. Shum, "Image segmentation by data driven markov chain monte carlo," ICCV, 2001.

[22] S.C. Zhu, R. Zhang, and Z.W. Tu, "Integrating top-down/bottom-up for object recognition by ddmcmc," *CVPR*, 2000.

[23] A. Yuille and D. Kersten, "Vision as bayesian inference: analysis by synthesis," *TRENDS in Cognitive Sciences*, vol. 10, no. 7, pp. 301–308, 2006.

[24] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," CVPR, 2008.

[25] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *PAMI*, vol. 6, pp. 721–741, 1984.

[26] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, 2009.

[27] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *CVPR*, 2009.

[28] L Gorelick, M Blank, E Shechtman, M Irani, and R Basri, "Actions as space-time shapes," *PAMI*, vol. 29, no. 12, pp. 2247–2253, 2007.

[29] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *ECCV*, 2010.

[30] David I. Hastie and Peter J. Green, "Model choice using reversible jump markov chain monte carlo," *Statistica Neerlandica*, vol. 66, no. 3, pp. 309–338, 2012.

[31] Zoltan Kato, "Segmentation of color images via reversible jump mcmc sampling," *Image Vision Comput.*, vol. 26, no. 3, pp. 361–371, Mar. 2008.

[32] Y. Deng, , and B. S.Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *PAMI*, vol. 23, no. 8, pp. 800–810, Aug 2001.

[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *ICCV*, 2001.