Fast Unsupervised Ego-Action Learning for First-Person Sports Videos

Kris M. Kitani
UEC Tokyo
Tokyo, Japan
kitani@is.uec.ac.jp

Takahiro Okabe, Yoichi Sato
University of Tokyo
Tokyo, Japan
takahiro,ysato@iis.u-tokyo.ac.jp

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

Abstract

*Portable high-quality sports cameras (*e.g. *head or helmet mounted) built for recording dynamic first-person video footage are becoming a common item among many sports enthusiasts. We address the novel task of discovering first-person action categories (which we call ego-actions) which can be useful for such tasks as video indexing and retrieval. In order to learn ego-action categories, we investigate the use of motion-based histograms and unsupervised learning algorithms to quickly cluster video content. Our approach assumes a completely unsupervised scenario, where labeled training videos are not available, videos are not pre-segmented and the number of ego-action categories are unknown. In our proposed framework we show that a stacked Dirichlet process mixture model can be used to automatically learn a motion histogram codebook and the set of ego-action categories. We quantitatively evaluate our approach on both in-house and public YouTube videos and demonstrate robust ego-action categorization across several sports genres. Comparative analysis shows that our approach outperforms other state-of-the-art topic models with respect to both classification accuracy and computational speed. Preliminary results indicate that on average, the categorical content of a 10 minute video sequence can be indexed in under 5 seconds.*

1. Introduction

Affordable rugged high-quality head-mounted cameras for recording an athletes first-person point-of-view experience is the newest digital toy for professionals and amateurs alike (Figure 1). In this paper, we present a novel approach for quickly indexing videos to enable efficient search for these types of first-person sports videos (Figure 2). We motivate this work with an example. Imagine that a mountain bike racer has just finished running several laps through a practice course and would like to review the third jump of his second lap before he proceeds with the rest of his training. How can he locate the correct position in the video?

Figure 1. GoPro camera used to generate first-person POV videos.

Figure 2. First-person point-of-view for various sports.

A typical manual search process would usually require the user to fast-forward through the video sequence to find the desired location. However, if the video was indexed by action categories, he could immediately review say, a color-coded time index (Figure 3 and 10) and go directly to the desired location in the video.

What kind of requirements would such functionality necessitate? An unsupervised approach would be best since access to labelled training data may be limited and we do not want to burden the user with a data labeling process. A supervised solution leveraging human computation may also be valid but would require an intensive training process. Near real-time processing is another requirement since we cannot expect the user to wait hours or even several minutes for results. Finally, the extracted video feature should be discriminative enough to differentiate between action categories but also robust enough to deal with extreme ego-motion.

Sports videos are usually characterized by extreme ego-motion that far exceeds the motion addressed in traditional ego-motion analysis (*e.g.* wheeled vehicles or robots). This intense camera motion translates to large frame-to-frame displacement, significant motion parallax, motion-blur and the rolling-shutter effect (*i.e.* video wobble). Methods that require accurate feature tracking or high-quality registration

Figure 3. Color-coded video time bar indexed by category can be used to find ego-actions of the same category.

Figure 4. Distortion examples: projectiles, motion blur, water.

are rendered infeasible in most situations.

Though the task of image-based analysis may seem daunting there are unique characteristics of first-person sports videos that can also be leveraged to our advantage. In particular, a video contains a single sport, performed with the same person and recorded with the same camera. This has three implications: (1) the list of possible actions is constrained by nature of the sport (*i.e*. certain actions are repeated throughout the video), (2) actions of the same category will look similar because they are performed by the same person (more so than across multiple actors) and (3) actions of the same category share similar image distortion (*e.g*. jumping causes a rolling shutter effect). Our approach leverages these conditions to categorize even the most extreme motions.

Furthermore, we know that human motion observed from a first-person point-of-view is global and not local. This means that we should be able to aggregate global motion and marginalize out local outlier motion. We also know that motion involving the human gait has an inherent frequency component. Therefore we can expect that frequency analysis can be used as a salient feature for ego-action categorization.

To the best of our knowledge, this is the first work to deal with the novel task of discovering *ego-action categories* from first-person sports videos. We propose the use of a simple global representation of motion that is both robust and discriminative. Our proposed approach shows the applicability of Dirichlet process mixture models to novel real-world problems and the powerful potential of online inference. We also provide a new labeled benchmark dataset for standardized analysis of dynamic outdoor first-person sports videos.

2. Related Work

Prior work on vision-based first-person human action analysis has been limited to indoor activities with a focus on hand gesture recognition. Early work with vision-based analysis for head-mounted cameras focused on hand gesture recognition for sign language recognition [10] and context aware gesture recognition [11]. In recent years, there has been a renewed interest in first-person vision with a similar focus on object recognition [7], hand gesture recognition [6, 13] and hand tracking [12]. These works are important because understanding indoor first-person activities has a clear social impact (*e.g*. patient monitoring, assistive technologies). We aim to expand the research domain by addressing dynamic outdoor activities which involve more full body motion.

Work with body worn sensors have also been shown to be effective in categorizing human action and activity categories. Inertial motion sensors have been used to discover long-term activities (*e.g*. working at the office, eating dinner) [4] and segmenting a signal into primitive units [9]. It has also been demonstrated that more complex activities can be learned using an ensemble of body worn sensors such as motion sensors, force sensing resistors and ultra-wide band tags [17]. While we expect that a comprehensive solution for first-person action analysis will require multi-modal sensor fusion, the goal of this work is to examine the extent (and limitations) of vision-based strategies for first-person action categorization.

We also note that our work is different from work on vision-based ego-motion estimation, visual odometry, structure-from-motion and autonomous navigation because we are concerned primarily with the task of ego-action categorization and not (necessarily) accurate motion estimation and localization.

3. Motion Feature Extraction

As we have stated earlier, first-person sports footage can be very noisy. We observe different types of image distortion such as motion blur, the rolling shutter effect (*i.e*. many sports cameras use CMOS sensors), motion parallax, and environmental factors such as splashing water, glare and projectiles (Figure 4). Therefore, we will need a representation of motion that is robust to a certain degree of image distortion.

We use sparse optical flow vectors as our basic motion feature. Although quick movement causes many false correspondences, we also observe that the general direction and relative magnitude of a subset of the flow vectors are usually consistent. We take advantage of this phenomenon and extract the set of internally consistent flow vectors by keeping only those points that can be accounted for by a planar homography (we use RANSAC). This step effectively re-

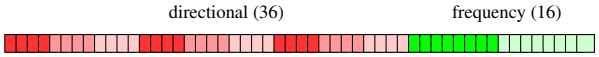

Figure 5. 52-dimensional motion histogram is a concatenation of 36 directional bins and 16 frequency bins.

moves flow vectors caused by bad matching or local motion caused by other moving objects in the scene. We add further robustness by aggregating the set of flow vectors as a normalized motion histogram.

While designing the motion histogram we found that there are two types of motion that are needed to discriminate between different ego-action categories (details discussed in section 7.3). The first component is instantaneous motion (directional component) and periodic motion (frequency component). For example, the action of *turning* one's head has a strong directional component, while repetitive actions like *walk* and *run* have strong periodic components.

The directional component of motion is encoded using a simple 36 bin histogram indexed over the quantized joint space of 4 flow directions, 3 flow magnitudes and 3 flow variance (aggregate difference of flow magnitudes against the average flow magnitude). The frequency component is computed with the discrete Fourier transform (DFT) over the average flow magnitude with a sliding window of 64 frames. The X (horizontal motion) and Y (vertical motion) components are analyzed separately. The frequency component is encoded with a histogram of 16 bins. The first 8 frequency amplitude components of the X channel and Y channel are thresholded, normalized and added to the first 8 bins and last 8 bins, respectively. The directional histogram and frequency histogram are concatenated to yield a motion histogram $\mathbf{y} = [y_1, \ldots, y_M]$ where $M = 52$ in our representation (Figure 5).

4. Dirichlet Process Mixture Models

Since we would like to infer the number of ego-action categories and the motion histogram codebook automatically from the data, we use a hierarchical Bayesian model which we call stacked Dirichlet Process Mixtures. We begin with a brief explanation of the Dirichlet Process (DP). Readers already familiar with Dirichlet process mixtures (DPM) can skip to Section 5.

4.1. Understanding the Dirichlet Process

To better conceptualize the Dirichlet process it is helpful to compare it to the finite Dirichlet distribution. A Dirichlet distribution is defined as

$$Dir(\boldsymbol{\pi}; \alpha_1, \ldots, \alpha_K) = B(\alpha_1, \ldots, \alpha_K)^{-1} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad (1)$$

where $\alpha_1, \ldots, \alpha_K$ are the non-negative parameters of the distribution, the normalization constant $B(\alpha_1, \ldots, \alpha_K)$ is the Beta function of the parameters, and $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ is a K-dimensional probability function where $\sum_k \pi_k = 1$.

A useful analogy is to interpret the Dirichlet distribution as a big urn containing many biased K-faced dice, where a single draw from the urn yields one biased K-faced die. For an arbitrary die drawn from the urn, the expected value of the probabilities over each of the K faces (outcomes) is defined as $E[\boldsymbol{\pi}] = \{\frac{\alpha_1}{\alpha_0}, \ldots, \frac{\alpha_K}{\alpha_0}\}$ where $\alpha_0 = \sum_k \alpha_k$. The expected value of the probability of a single face (outcome) is $E[\pi_k] = \{\frac{\alpha_k}{\alpha_0}\}$. Alternatively, the vector $\boldsymbol{\mu} = \{\frac{\alpha_1}{\alpha_0}, \ldots, \frac{\alpha_K}{\alpha_0}\}$ can be used to write the Dirichlet distribution as $Dir(\boldsymbol{\pi}; \alpha_0 \boldsymbol{\mu})$. Next we show how this alternative parameterization is related to the DP.

The Dirichlet process can be understood as a generalization of the Dirichlet distribution with infinite K.

$$DP(\boldsymbol{\pi}; \alpha_0, \boldsymbol{\mu}) = \lim_{K \to \infty} \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad (2)$$

The parameter α_0 is typically called the concentration parameter and $\boldsymbol{\mu}$ is called the base distribution (where $\boldsymbol{\mu}$ is continuous in the general case). The expected value of the probability of a given outcome is defined as $E[\pi_k] = \frac{\alpha_k}{\alpha_0}$ or $E[\boldsymbol{\pi}] = \boldsymbol{\mu}$ in vector form.

The Dirichlet distribution (and the DP) is a useful prior distribution when one is working with K dimensional histograms. When a histogram is interpreted to be generated from multiple draws from a probability function $\boldsymbol{\pi}$, the likelihood of the histogram takes the form of a product of probabilities (formally called a *multinomial distribution*). When the multinomial distribution is multiplied with a Dirichlet distribution (also a product of probabilities), the result is another Dirichlet distribution. Formally, this result is due to the fact that the Dirichlet distribution is the *conjugate prior* of the multinomial distribution. We will see how this property is used in Section 6.1.

4.2. Chinese Restaurant Process

An alternative perspective used to explain the DP is the Chinese restaurant process (CRP). A CRP describes the process of generating a probability distribution $\boldsymbol{\pi}$ from a DP. Using the analogy of customers arriving at a Chinese restaurant, the CRP seats (assigns) the newest d-th customer to a table k, denoted as z_{dk} (shorthand for $z_d = k$), according to the following probability distribution.

$$p(z_{dk} | z_1, \ldots, z_{d-1}; \alpha_0) = \begin{cases} \frac{c(k)}{\alpha_0 + d}, & k \leq K_d \\ \frac{\alpha_0}{\alpha_0 + d}, & k > K_d \end{cases} \quad (3)$$

where z_i is the table assignment of the i-th customer and $c(k)$ is the current number of points assigned to cluster k and K_d is the number of occupied tables. The probability of being seated at a certain table and the number of occupied tables K_d depends on how the first $d-1$ customers

have been seated. The important result to observe is that the CRP (equivalently the DP) generates a (potentially infinite) discrete probability distribution $\boldsymbol{\pi}$. Again, we will see how this representation is utilized in section 6.1.

4.3. Dirichlet Process as a Mixture Prior

A typical mixture model is composed of an observation $\boldsymbol{x}$ and a (latent) topic $\boldsymbol{z}$, where the model is specified by the observation likelihood $p(\boldsymbol{x}|\boldsymbol{z})$ and the prior over mixtures (also called topics or categories) $p(\boldsymbol{z})$. In a discrete parametric model, $p(\boldsymbol{z})$ is a discrete K valued probability function. In a Bayesian Dirichlet process mixture model, the prior distribution over the mixture distribution $\boldsymbol{\pi} \triangleq p(\boldsymbol{z})$ is the Dirichlet Process $DP(\boldsymbol{\pi}; \alpha_0, \boldsymbol{\mu})$.

We are most interested in the use of DPMs to discover the number of mixture components from data. Since a proper treatment of the DP involves computing complex integrals over the infinite DP, in practice, approximate inference algorithms are used to estimate the number of mixtures. Typical inference algorithm include MCMC [5], variational inference [1], multi-pass online inference [15] and online beam search [2].

5. Stacked Dirichlet Process Mixtures

In our framework we take a sequential bottom-up approach, in which we learn the motion codebook with a single DPM and then pass the results to a second DPM to learn the ego-action categories. This stacked architecture is similar to a hierarchical Dirichlet process mixture model in that topics are organized in a hierarchy and all topics (i.e. ego-actions and codewords) at every level in the hierarchy are constrained to share the same discrete set of possible observations (*i.e.* motion histograms). The two model topologies are shown in Figure 6. The use of hierarchical mixture models have been shown useful modeling hierarchical grouping in such tasks as surveillance videos [16], documents analysis [14] and object categorization [8].

The main advantage of the stacked DPM is that it allows us to decouple the inference over a hierarchical structure into simpler inference over two DPMs. We show in the next section how this decoupling allows us to run very fast online inference.

6. Task of Inference

First, a single video is cut into D equally sized video splices $d \in D$, where each video splice d is made up of N_d frames (N_d=60 in our experiments) and each frame is indexed by $n \in N_d$. The input to our model is a set of motion histograms $\mathbf{Y} = \{\boldsymbol{y}_{11}, \cdots, \boldsymbol{y}_{nd}, \cdots, \boldsymbol{y}_{N_D D}\}$ generated from each frame n in the video. As the output, our model returns the ego-action index (cluster assignments) $\mathbf{Z} = \{z_1, \cdots, z_d, \cdots, z_D\}$, where z_d is an indicator variable that contains the ego-action cluster assignment for the d-th video splice in the video. The total number of ego-action clusters K is estimated as part of the inference process.

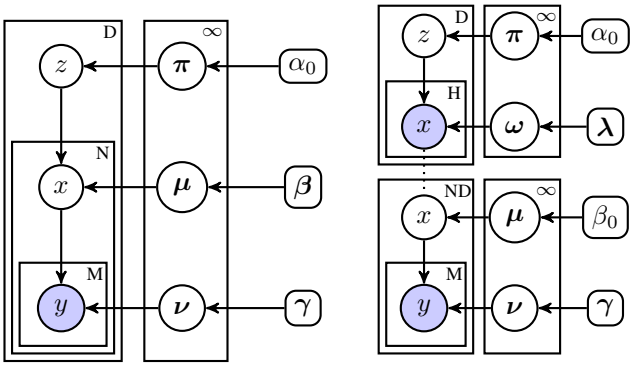

Figure 6. (a) Hierarchical Dirichlet Process Multinomial Mixture, (b) Stacked Dirichlet Process Multinomial Mixture. Graphs drawn using explicit cluster indicators and stick-breaking representation.

The entire inference process is composed of two unsupervised learning steps. In the first step, the motion codebook is learned and in the second step, the ego-action categories are discovered. What is unique about our approach is that we efficiently learn the codebook and accumulate a histogram over the codewords for each video slice in a single pass over the data. We are enabled to do this by online inference [15] which we outline below.

6.1. Inferring the Motion Codebook

The input into the first DPM is a set of motion histograms $\mathbf{Y}$ which are processed sequentially. The output of the first DPM is a sequence of codeword assignments $\mathbf{X} = \{x_{11}, \cdots, x_{nd}, \cdots, x_{N_d D}\}$, where each indicator variable x_{ndh} (shorthand for $x_{nd} = h$) has been assigned to the codeword h that maximizes the following posterior probability.

$$\hat{h} = \arg\max_h \left\{ p(x_{ndh} | \boldsymbol{y}_{nd}, \mathbf{Y}^{-nd}, \mathbf{X}^{-nd}) \right\} \quad (4)$$

$$\propto \arg\max_h \left\{ p(x_{ndh} | \mathbf{X}^{-nd}) p(\boldsymbol{y}_{nd} | \mathbf{Y}^{-nd}, \mathbf{X}^{nd}) \right\} \quad (5)$$

Here $\mathbf{Y}^{-nd}$ is the set of all past observations (motion histograms) excluding $\boldsymbol{y}_{nd}$ and $\mathbf{X}^{-nd}$ is the set of all past codeword assignments excluding x_{nd}. The hyperparameters (DP concentration α_0, Dirichlet parameters β_0) have been omitted from the notation for simplicity. Using Bayes rule and dropping variables due to conditional independence, this posterior can be decomposed (Eq. 5) into two terms: (1) the current prior over cluster assignments and (2) the likelihood of the observed motion histogram.

The first term is the DP prior probability of a cluster assignment and is modeled with a mixture of DPs with L dif-

ferent concentration parameters.

$$p(x_{nd}|\mathbf{X}^{-nd}) = \sum_{l}^{L} p(\alpha_{0l}) p(x_{nd}|\mathbf{X}^{-nd}; \alpha_{0l}) \quad (6)$$

where each DP indexed by l is modeled using a CRP.

This allows the DP concentration parameter to be free and the weights $p(\alpha_{0l})$ over the CRPs can be inferred online with the following update.

$$\hat{p}(\alpha_{0l}) = p(\alpha_{0l}) p(x_{nd}|\mathbf{X}^{-nd}; \alpha_{0l}) \quad (7)$$

The second term is the motion histogram likelihood and must be computed by integrating over the prior conditional Dirichlet distribution q over the probability distribution $\boldsymbol{\nu}_h = \{\nu_{h1}, \ldots, \nu_{hM}\}$.

$$\begin{aligned} \mathcal{L} &= p(\boldsymbol{y}_{nd}|\mathbf{Y}^{-nd}, \mathbf{X}^{-nd}, x_{ndh}) & (8) \\ &= \int_{\nu_h} p(\boldsymbol{y}_{nd}|\boldsymbol{\nu}_h) q(\boldsymbol{\nu}_h|\mathbf{Y}^{-nd}, \mathbf{X}^{-nd}, x_{ndh}) d\boldsymbol{\nu}_h & (9) \\ &= E_q[p(\boldsymbol{y}_{nd}|\boldsymbol{\nu}_h)] & (10) \end{aligned}$$

The first term inside the integral of equation (9) is the multinomial distribution $p(\boldsymbol{y}_{nd}|\boldsymbol{\nu}_h) = \prod_m^M p(\nu_{hm})^{y_{mnd}}$ and it describes the likelihood of multiple draws from the discrete distribution $\boldsymbol{\nu}_h$ of the h^{th} codeword. The second term inside the integral is the conditional Dirichlet distribution q, which is a product of the multinomial distribution of all past observations and a base Dirichlet distribution.

Fortunately, the final form of the likelihood $\mathcal{L}$ conveniently reduces to a product of simple fractions,

$$E_q[p(\boldsymbol{y}_{nd}|\boldsymbol{\nu}_h)] \propto \prod_m^M \left[\frac{c(h,m) + \beta_0/M}{\sum_{m'} c(h,m') + \beta_0} \right]^{y_{mnd}} \quad (11)$$

where $c(h,m)$ is the total count of flow vectors in bin m accumulated from all motion histograms that belong to codeword h and y_{mnd} is the count of the m^{th} bin in the motion histogram $\boldsymbol{y}_{nd}$. This is because the Dirichlet distribution is the conjugate prior of the multinomial distribution (recall a histogram follows a multinomial distribution) and the conditional expected value of the Dirichlet parameters can be computed as the ratios of the empirical evidence $c(h,m)$ plus the concentration hyper-parameter β_0 (which is set to 1 for all experiments).

Since this inference algorithm only adds new codewords to the codebook (*i.e.* never removes codewords), the order of the codewords is always preserved. This property allows us to accumulate the histogram of codewords $\boldsymbol{x}'_d$ in a single pass by keeping a record of the counts over codewords for each video splice d.

We also note that while the CRP does allow for the size of the codebook to be infinitely large, the size of the codebook is in practice log-bounded (Figure 7).

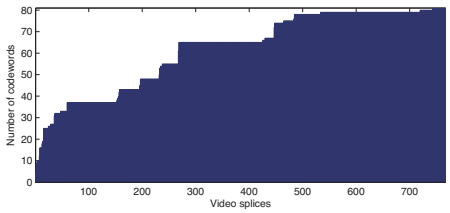

Figure 7. Logarithmic growth of the codebook for the PARK sequence.

6.2. Inferring Ego-action Categories

The input to the second DPM is a set of histograms over codewords, $\boldsymbol{x}'_d = \{x'_1, \ldots, x'_H\}$, for every video splice $d \in D$, where H is the total number of motion codewords. The output is a set of assignments $\mathbf{Z} = \{z_1, \cdots, z_d, \cdots, z_D\}$, where each assignment variable z_{dk} (shorthand for $z_d = k$) maximizes the posterior distribution.

$$\hat{k} = \arg\max_k \left\{ p(z_{dk}|\boldsymbol{x}'_d, \mathbf{X}'^{-d}, \mathbf{Z}^{-d}) \right\} \quad (12)$$

where $\mathbf{X}'^{-d}$ is the set of all previous observations (codeword histograms) excluding $\boldsymbol{x}'_d$ and $\mathbf{Z}^{-d}$ is the set of all previous clusters (ego-action category) assignments excluding z_d. The posterior is in the same form as equation (4) and is decomposed in the same way as equation (5).

In the first quantization step, inference was performed at every frame. In the second layer, we now run inference over several permutations of the observed motion histograms, to search for a more optimal clustering. We can identify a good ordering $\hat{\boldsymbol{r}}$ (a D dimensional vector dictating the order of the data) that maximizes $\arg\max_{\boldsymbol{r}} \{\tilde{\mathcal{L}}_{\boldsymbol{r}}\}$, where $\tilde{\mathcal{L}}$ is the pseudo marginal likelihood [3].

$$\begin{aligned} \tilde{\mathcal{L}}_{\boldsymbol{r}} &= \prod_d^D p(x_d|\mathbf{X}'^{-d}_{\boldsymbol{r}}, \mathbf{Z}^{-d}_{\boldsymbol{r}}) & (13) \\ &= \prod_d^D \sum_k^K p(z_{dk}|\mathbf{X}'^{-d}_{\boldsymbol{r}}, \mathbf{Z}^{-d}_{\boldsymbol{r}}) \\ &\times \int p(\boldsymbol{x}'_d|\boldsymbol{\omega}_k) q(\boldsymbol{\omega}_k|\mathbf{X}'^{-d}_{\boldsymbol{r}}, \mathbf{Z}^{-d}_{\boldsymbol{r}}) d\boldsymbol{\omega}_k & (14) \\ &\approx \prod_d^D \sum_k^K p(z_{dk}|\mathbf{X}'^{-d}_{\boldsymbol{r}}, \mathbf{Z}^{-d}_{\boldsymbol{r}}) E_q[p(\boldsymbol{x}'_d|\boldsymbol{\omega}_k)] & (15) \end{aligned}$$

Again, the integral over the parameters $\boldsymbol{\omega}_k$ of the Dirichlet distribution associated with the k^{th} ego-action category reduces to the same fractional form as equation (11). To speed up computation $q(\boldsymbol{\omega}_k|\mathbf{X}'_r, \mathbf{Z}_r)$ is used in place of $q(\boldsymbol{\omega}_k|\mathbf{X}'^{-d}_r, \mathbf{Z}^{-d}_r)$. We note that other strategies like beam search [2] can be implemented to improve over this naive search strategy. However, as we will show in the next section, we found the random search strategy to be quite sufficient.

Biking (BIK) Surfing (SRF) Skiing (SKI) Slope style (SLO) Snowboarding (SNO) Horseback riding (HOR)

Figure 8. YouTube sports dataset.

7. Discovering Ego-action Categories

7.1. Evaluation Videos

To evaluate the performance of our method we need both choreographed video sequences to run controlled experiments and a broad span of real-world videos to observe the performance of our method across several sport genres.

We choreographed two videos. The first video (QUAD) consists of 124 video splices (a video splice contains 60 frames) and contains 11 different ego-actions. This video was primarily used to show how the design of our motion features effect performance. The second video (PARK) is a 25 minute workout video which contains 766 video splices and contains 29 different ego-action categories. The PARK sequence was used to investigate the performance of our method over a larger set of ego-action categories.

We also utilized six real-world first-person YouTube sports videos (images in Figure 8) from mountain biking (BIK), longboard surfing (SRF), skiing (SKI), slope-style (SLO), snowboarding (SNO) and horseback riding (HOR) (details in Table2). These publicly available datasets show the performance of our method over multiple sports genres. All sequences used for evaluation were recorded with the GoPro HD camera (Figure 1) and the ego-action category names for all sequences are given in Table 1.

Table 1. Ego-action category names.

QUAD	jump, run, leftstand, rightupdown,updown, stand, rightleft, runturnleft, standturnright, walk, walkinplace
PARK	upright, jog, updown, downright, downforward, slowjog, run, twist, walk, pullups, downrightleft, crawnet, downleft, down, right, stop, ditdown, up, slowwalk, pivotright,walkleft, standup, lookdown, left, exit, slide, reachup, hold, rest
BIK	quicklookleft, bumpyright, fastright, curveleft, straight, rough, cruise, left, straightright, jump, highjump
SRF	underwater, backrightleft, left, hitwave, forwardback, situnderwater, backright, lightpaddle, standup, forward, sitstop, kickpaddle
SKI	rotateleftdown, lookup, downright, hopdown, turnrightleft, turnright, slowdownleft, upturnright, leftright, turnleft, turnleftright, slowdown, smoothright, wedgeleft, hopturnleft
SLO	highfive, spin, liftspindown, left360, pedal, land, rotateright, leftrightshake, straight, updownshake, leftloopup, inplane360, leftdown, rampup
SNO	lookdownright, lookrightrotate, rightdownright, hardfrontedge, shakefalldown, lookup, shakelookleft, standup, rightface, fallslide, rotateright, bumpy, rotateleft, upshake, shake, shakeright, still, forwardbackedge, lookleftright, hardbackedge, lookrightleft, leanforward, hitbump, backedge, lookright, forward
HOR	intro, jump, speedup, landing, startrun, prepjump, pulltrot, yanktrot, jumpland, slowtrot, straight, slowright, landjump, lowjump, bumptrot

Table 2. YouTube sport video details and performance.

	BIK	SRF	SKI	SLO	SNO	HOR
Video splices	104	26	34	35	77	100
Num. ego-actions	24	12	18	15	26	15
Avg. F-measure	0.54	0.64	0.94	0.47	0.67	0.49
Comp. time (s)	0.44	0.10	0.18	0.45	0.57	0.35

7.2. Matching Categories for Analysis

We use the F-measure (F) to measure performance. The F-measure is the harmonic mean of the precision P and recall R, defined as $F = 2PR/(P + R)$. Values range from 0 to 1, where 1 represents perfect performance. To compute the average F-measure we must find the one-to-one correspondence between the ground truth labels to discovered ego-action categories. We perform a greedy search by identifying the best match (*i.e.* best F-measure) between a single ground truth label and a discovered ego-action category. We remove that pair as a possible future candidate, and repeat the process for the remaining ground truth labels and discovered ego-action categories. When the number of discovered ego-action categories is not equal to the number of ground truth categories, the extra categories are appended to the list with an F-measure of zero. The average F-measure is computed from a weighted average of the F-measures of each category. Categories with no correspondences always lower the average F-measure.

7.3. Feature Design

As stated earlier, each flow vector is binned depending on the flow direction, flow magnitude and flow variance. Flow direction was quantized into 4 directional bins (left, right, up, down) although tests showed similar performance for finer bins. Magnitude was quantized into three bins in increments of 8, which were determined from performance on the QUAD sequence. The bins roughly divide motion magnitudes into small (1-8), medium (8-16) and large (16+). In addition to direction and magnitude, we observed that the variance of the magnitude of flow vectors for a single frame was also an informative feature for dis-

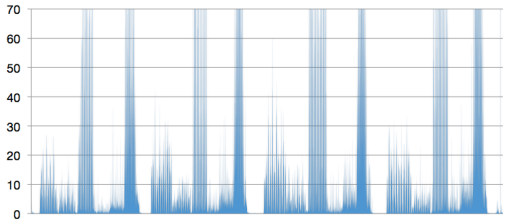

Figure 9. Peaks in flow **variance** produced by actions with extreme changes in acceleration caused by floor impact (*e.g.* jump). Horizontal axis is time and the vertical axis is the variance. The eight peaks are generated by the four repetitions of *jump* and *run*.

criminating between ego-actions. The variance bins divide motion magnitude variance into small (0-9), medium (9-18) and large (18+). Figure 9 shows how variance peaks for ego-actions with large changes in acceleration such as *run* and *jump*. For the QUAD sequence, using only the directional component gave an average F-measure score of 0.83 and using only the frequency component yielded 0.72. The joint use of both components returned a score of 0.93. This trend was found to be true over all of our videos.

Based on the observation that most actions have a single dominating axis led us to represent periodic human motion using frequency analysis on the horizontal axis and vertical axis independently. The amplitude of the top eight frequencies for each axis were treated as bins to capture dominant periodic motion. Each channel was thresholded using the running average of that channel and absolute thresholds for x and y, $t_x = 50$ and $t_y = 200$ respectively. When the cumulative sum of the frequency bins passed 500, the frequency channels were normalized to sum to 500. Similar results were obtained for normalization values between 250 and 1000.

7.4. Performance Across Sports Genres

For all videos, we labelled every video splice (2 second segment) by visual inspection to generate the ground truth. Then discovered labels were compared against the ground truth using the F-measure. A visualization of the performance across the videos are shown in the form of matching matrices in Figure 11.

As expected, we perform better on choreographed videos because the ego-actions categories were known in advance and ego-actions occurred in sequence over long durations. The average F-measure performance was 0.93 and 0.72, for the QUAD and PARK sequence, respectively. Although the PARK sequence is significantly longer than the QUAD sequence and contains almost three times as many actions, we get relatively good performance. An interesting observation we made was that certain actions, for example *jog*, change over time (*i.e.* the athlete jogs slower toward the end of the 25 minute work out). This makes sense since we expect that the athlete will become more fatigued toward the end of the workout. However, in our current model we are not

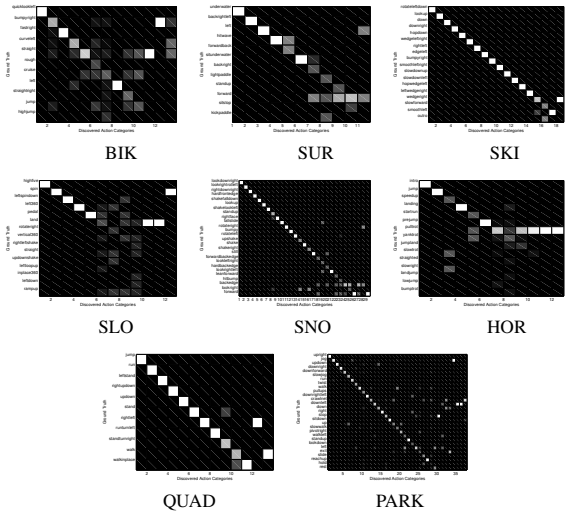

Figure 11. Matching matrix visualization of performance across sports genres. Vertical axis is the ground truth label and the horizontal axis is the discovered ego-action categories. Perfect performance yields an identity matrix.

able to adapt these types of changes over time.

Across the sports genres, skiing, surfing and snowboarding yielded higher scores over an average F-measure of 0.6, while horseback riding, mountain bike, slope-style scored below 0.6 (see Table 2). We attribute this difference in performance to the fact that the former sports genres contain strong periodic ego-actions while, the latter genres contain less periodic signals, with the exception of horseback riding. Although we expected horseback riding to produce more salient periodic ego-actions, we found that the proximity of the horse and the riders hands to the camera had an adverse effect on the optical flow calculations. We also attribute the difficulty with the bicycle sequences to symmetric motion fields (*e.g.* zoom and in-plane rotations) generated by a fast moving (or spinning) bicycle. Our current representation can not differentiate between symmetric motion fields.

7.5. Comparative Evaluation

We compare the top-level DPM with online inference (DPM-OL) against the DPM with variational inference (DPM-VI), latent Dirichlet allocation with variational inference and sequential importance sampling (LDA-VI), probabilistic latent semantic analysis with the EM algorithm (PLSA-EM), a naive Bayesian mixture model with the EM algorithm (NBM-EM) and K-means clustering. The maximum number of iterations was set to 100 and the stop criteria was set to 10^{-5} (change in the log-likelihood). Each algorithm was randomly initialized 20 times, except for DPM-OL and K-means which were initialized 200 times. The true number of actions were given to all models except for the two DP models.

Despite the fact that we provide the true K for the non-

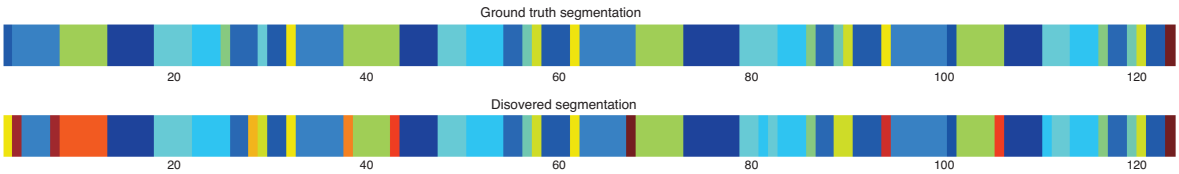

Figure 10. Color coded ego-action category index for the QUAD sequence discovered with our method. Horizontal axis is time and colors represent different ego-action categories.

Table 3. Detailed comparison for choreographed datasets.

QUAD	F-measure	P	R	K	sec.
DPM-OL	**0.93**	**0.95**	**0.92**	13	0.47
DPM-VI	0.92	0.94	0.92	12	10.12
LDA-VI	0.87	0.89	0.87	11	3.38
PLSA-EM	0.89	0.91	0.89	11	2.88
NBM-EM	0.66	0.59	0.91	5	**0.25**
K-means	0.89	0.89	0.91	9	1.44

PARK	F-measure	P	R	K	sec.
DPM-OL	**0.71**	**0.76**	0.71	37	8.69
DPM-VI	0.61	0.66	0.62	40	73.64
LDA-VI	0.56	0.56	0.66	27	30.99
PLSA-EM	0.53	0.58	0.59	29	63.51
NBM-EM	0.44	0.38	**0.73**	10	**3.75**
K-means	0.53	0.62	0.52	29	25.04

DP models, we observe that the online inference DPM-OL performs the best and works surprisingly well (Table 3). Interestingly, results show that non-DP models tend to underestimate the true K. Also, alternating algorithms maximizing non-convex functions seem to be more sensitive to initialization values and require many trials to find a good starting point.

7.5.1 Computation Time

The wall-clock computation times for our comparative experiments are given in Table 3. Our approach ranks second in speed next to the EM algorithm with the naive Bayes mixture model (NBM-EM). Taking the average computation time across datasets, 2 minutes of video (60 segments) can be processed in less than a second, with our method. Variational inference with the DPM takes significantly more time than the other models because the dimensionality of the variational distribution needs to be sufficiently large (*e.g.* 100 for our experiments). All experiments were performed with a 2.66 GHz CPU.

8. Conclusion

In this paper we have introduced the novel task of discovering *ego-action categories* from first-person sports videos. We have described the power of the Dirichlet process to infer motion codebooks and ego-action categories with no training data. Furthermore, we have shown that DPMs can be applied to difficult real-world problems without incurring the cost of computational complexity. In particular, we have shown that online inference can perform on par with other types of approximate inference over topic models, while also providing a significant savings in computational cost. Our preliminary experiments suggest that vision-based ego-action analysis can be successfully applied to dynamic first-person videos.

References

[1] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2006. 3244

[2] H. Daume. Fast search for Dirichlet process mixture models. In *AISTATS*, 2007. 3244, 3245

[3] S. Geisser and W. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979. 3245

[4] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *International Conference on Ubiquitous Computing*, pages 10–19, 2008. 3242

[5] S. Jain and R. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007. 3244

[6] W. W. Mayol and D. W. Murray. Wearable hand activity recognition for event summarization. In *ISWC*, pages 122–129, 2005. 3242

[7] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, pages 3137–3144, 2010. 3242

[8] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, pages 1–8, 2008. 3244

[9] E. H. Spriggs, F. D. la Torre Frade, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Workshop on Egocentric Vision*, 2009. 3242

[10] T. Starner, A. Pentland, and J. Weaver. Real-time American sign language recognition using desk and wearable computer based video. *PAMI*, 20(12):1371–1375, 1998. 3242

[11] T. Starner, B. Schiele, and A. Pentland. Visual context awareness via wearable computing. In *ISWC*, pages 50–57, 1998. 3242

[12] L. Sun, U. Klank, and M. Beetz. EYEWATCHME: 3D hand and object tracking for inside out activity analysis. In *Workshop on Egocentric Vision*, pages 9–16, 2009. 3242

[13] S. Sundaram and W. Cuevas. High level activity recognition using low resolution wearable vision. In *Workshop on Egocentric Vision*, pages 25–32, 2009. 3242

[14] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 3244

[15] L. Wang and D. Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, (0):1–21, 2010. 3244

[16] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical Bayesian models. In *CVPR*, pages 1–8, 2007. 3244

[17] A. Zinnen, C. Wojek, and B. Schiele. Multi activity recognition based on bodymodel-derived primitives. *Location and Context Awareness*, pages 1–18, 2009. 3242