Action Recognition Using Ballistic Dynamics

Shiv N. Vitaladevuni
Janelia Farm Research Campus,
Howard Hughes Medical Institute,
Ashburn VA 20147, U.S.A.
vitaladevunis@janelia.hhmi.org

Vili Kellokumpu
Machine Vision Group,
Univ. of Oulu,
Oulu, Finland.
kello@ee.oulu.fi

Larry S. Davis
Computer Vision Lab.,
Univ. of Maryland,
College Park, MD, U.S.A.
lsd@cs.umd.edu

Abstract

We present a Bayesian framework for action recognition through ballistic dynamics. Psycho-kinesiological studies indicate that ballistic movements form the natural units for human movement planning. The framework leads to an efficient and robust algorithm for temporally segmenting videos into atomic movements. Individual movements are annotated with person-centric morphological labels called ballistic verbs. This is tested on a dataset of interactive movements, achieving high recognition rates. The approach is also applied on a gesture recognition task, improving a previously reported recognition rate from 84% *to* 92%*. Consideration of ballistic dynamics enhances the performance of the popular Motion History Image feature. We also illustrate the approach's general utility on real-world videos. Experiments indicate that the method is robust to view, style and appearance variations.*

1. Introduction

We present a Bayesian framework for action recognition based on psycho-kinesiological observations, namely, the ballistic nature of human movements. Consider an everyday scenario, such as a person boiling water for brewing tea. This activity would consist of a number of actions such as reaching for a pot in the cupboard, putting it on the burner, etc. A typical adult familiar with the kitchen's layout would execute the movements efficiently, with rapid coordinated motion. Psycho-kinesiological studies have led to the following conclusions regarding such movements:

1. **Impulsive propulsion:** Ballistic movements are rapid and efficient, consisting of acceleration followed by deceleration [17].
2. **Simple trajectories, usually no mid-course correction:** The high speeds and impulsive propulsion result in relatively simple trajectories closely resembling straight lines and shallow 3D arcs [6, 18, 5, 9]. Empirical studies indicate that adults are very adept and efficient at planning ballistic movements [17]. Most movements are completed without any mid-course correction [6, 18]. Hesitations are rare and are not considered in this paper.
3. **Synchronized movement:** Psycho-kinesiological observations of reach and strike movements indicate that the hand's motion provides most of the dynamical information. Studies of reach movements involving torso rotation and stepping movements show that the entire body moves in synchrony with the hands with highly correlated velocities [9, 5].
4. **Inertial reference frame:** Humans plan their movements in an inertial reference frame, fixed during motion [9].

Due to their impulsive nature, these movements are referred to as "Ballistic". They form a vast majority of human interactive actions, evidenced by the large number of psychological studies, e.g. [17, 6, 18, 9, 5]. Ballistic movements are the natural unit in which humans plan everyday actions such as reach-to-grasp, pointing, punching, throwing, dancing and sports. Consider the following scenarios:

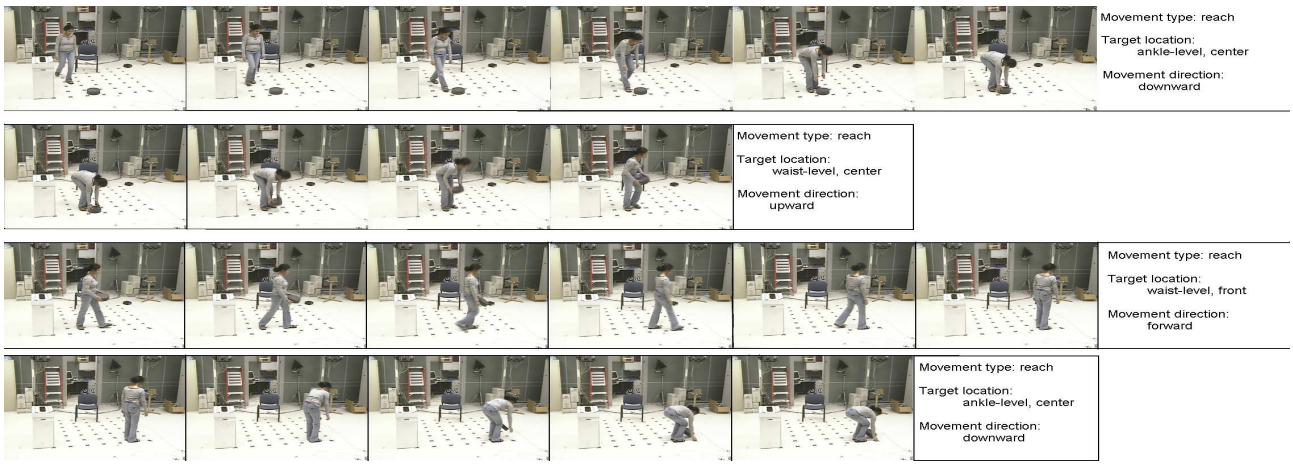
Figure 1(a): A person moves an object on the floor through a sequence of four ballistic movements - reaching down to grasp it, pick it up, step over to another location and reach down to place the object on the floor. The sequence was automatically segmented and annotated by our approach. Notice that the reach movements have different hand-target locations and the pose relative to the camera is variable.

Figure 1(b): Swing detection results on a tennis video taken from [4]. The two rows are key-frames for two instances of forehand swings showing the boundaries of the detected ballistic segments. The segments correspond to poising for the shot, retracting the hand, hitting the ball and following through. The video had 5 forehand swings - all are detected correctly.

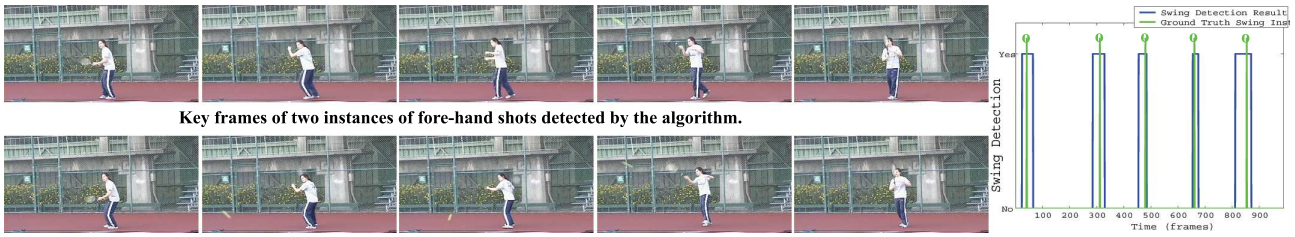
Figure 1(c): Key-frames at the boundaries of ballistic segments computed for a tutorial video of a dance called "Grapevine Pop". The tutor performs four cycles of the dance steps. The algorithm computes consistent segments in spite of the complicated, multi-limbed movements.

Figure 1(d): Three arm gestures used for signalling army vehicles. The key-frames defining boundaries of the ballistic segments are shown along with the segments' Motion History Images (MHIs) [2].

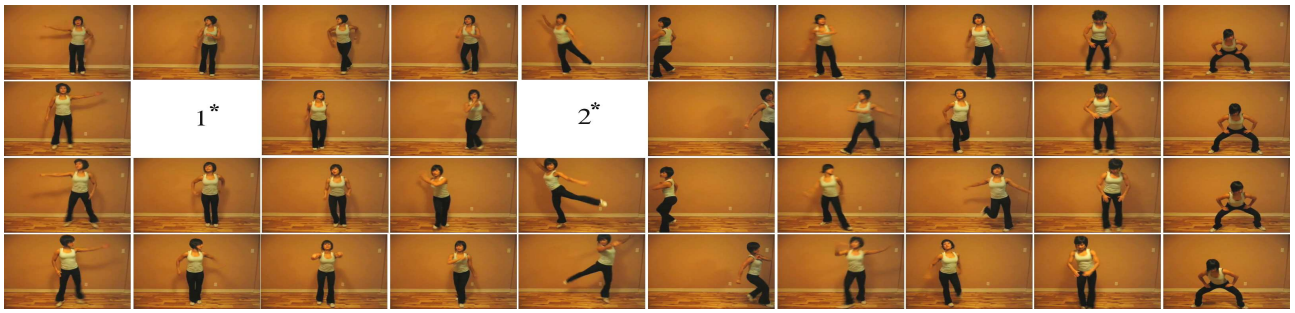
978-1-4244-2243-2/08/$25.00 ©2008 IEEE



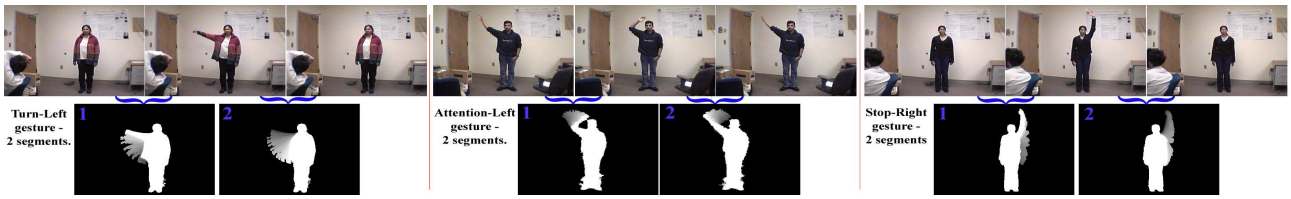
(a) Moving an object on the floor - reach to grasp the object, pick it up, walk to another place, place it down. The four movements are automatically segmented and labelled according to the hand's target location. Each row shows one segment, only every third frame is shown. The four reaches have different target locations, and the two "reach down" movements have different poses relative to the camera.



(b) Tennis forehand swing detection from video in [4]. The video had 5 forehands, all were correctly detected. The two rows show key-frames marking boundaries of ballistic segments for two forehand shots. The segments correspond to: poise to hit, retract hand, hit ball, and follow through. The plot on right shows the ground truth times of hitting the ball and the segments labelled as strikes. (See supplementary material.)



(c) Ballistic segments of a dance clip teaching "Grapewine Pop", curtesy www.fitmoves.com, duration 27 sec./810 frames. The tutor performs four cycles of dance steps, first to her right, then a symmetric set to her left, then again to her right and then another to her left. Each row corresponds to one cycle, the key-frames are synchronized based on the movements. The consistency of the segmentation across cycles indicates that ballistic dynamics forms a good unit for the movements. Two segment boundaries were missed - marked with numbers. In the case of the 1^{st} miss, the tutor was making small movements, like jogging in one place. In the 2^{nd} missed segment boundary, she almost jumped out of the camera view, causing erroneous optical flow. (See supplementary material.)



(d) Examples of three gestures used for army vehicle signalling and the corresponding ballistic segments. The second row shows Motion History Images for the segments. We present gesture recognition results using MHIs of ballistic segments. (See supplementary material.)

Figure 1. Ballistic movement scenarios: interactive movements, sports, dance and gestures. (Best viewed in color.)

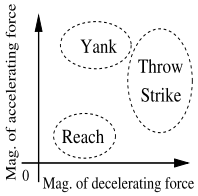


Figure 2. Varying the parameters of the ballistic movement model produces different types of movements: low acceleration and deceleration for reach, high acceleration and deceleration for throws and strikes, and high acceleration for yanking.

Following recent trends in vision, our approach is designed to work on single camera video, is robust to appearance and view variation, and does not require tracking bodyparts. The Bayesian framework leads to a robust and efficient algorithm for temporally segmenting videos into ballistic movements. Experiments with a dataset of 135 reaching, hitting and throw instances show it correctly segments more than 96% of the movements. All the segmentations in this paper, including the ones in Figure 1, were computed by the same model. After temporal segmentation, recognition is performed at the level of individual as well as sequences of movements. Individual movements are annotated with labels, called *ballistic verbs*, that describe the manner of propulsion - reaching or striking, and the target and direction of movement. The labels are person-centric and morphological, providing a natural basis for several applications such as video indexing and activity analysis. See Figures 1(a), 7, 8, and 1(b). Experiments indicate that the framework recognizes and labels more than 84% of a dataset of 135 reach and strike movements. Comparative experiments with alternative approaches indicate the approach's relative advantage. We also present an approach for recognizing complex actions through sequences of ballistic movements. This is illustrated with gesture recognition on a army signalling dataset - improving a previously published recognition rate of 84% [16] to 92%. Tests also indicate that the performance of MHI features is enhanced by considering ballistic dynamics.

2. Bayesian Model for Ballistic Dynamics

Psychologists have proposed two models for limb propulsion [17]: ballistic and mass-spring models, which form two ends of a spectrum of human movements. Ballistic movements involve impulsive propulsion of the limbs. There is an initial impulse accelerating the hand/foot towards a target, followed by a decelerating impulse to stop the movement. Reaching, striking, kicking are characteristically ballistic movements [17]. In the mass-spring model, the limb is modelled as a mass connected to springs (the muscles). The actuating force is applied over a period of time rather than impulsively [17]. Steady pushing, pulling, and many communicative gestures fall in this category. *Tai Chi* is also an example of mass-spring movements!

Varying the magnitudes of acceleration and deceleration produces different types of ballistic movements. Actions involving low acceleration and deceleration are said to exhibit reach dynamics [17]. These include reach-to-grasp, pointing, picking, etc. In contrast, actions involving high speeds and impacting force have high acceleration and deceleration. These are said to have strike dynamics, and include actions such as hitting, slamming and throwing. There is also the possibility of yanking in which the initial acceleration is high. See Figure 2 for an illustration. The movements may be further fine tuned according to the task at hand. E.g., when reaching for a small or fragile object, the deceleration phase sets in early and is prolonged to allow precise homing onto the target and grip formation [17].

The psycho-kinesiological observations stated in the introduction have useful implications for visual recognition, which form the basis for the Bayesian recognition model. These are tabulated in Figure 3.

Layer I of the Bayesian model corresponds to the sequence of movements composing an activity, executed to interact with objects and the environment. Each movement is considered to be independent of past and future movements given the context provided by the activity, and the states of the subject at the start and end of the movement. The equivalent Bayes net is shown in Layer I of Figure 4. Ballistic movements such as reaches and strikes are atomic in nature. Thus, the independence assumption is well suited for recognizing them.

Layer II in the Bayes net corresponds to the dynamics. A number of models have been proposed to analyze the trajectories of ballistic movements including the Minimum Jerk Model (MJM) [6] and Minimum Torque Change Model (MTCM) [18]. We use MJM for its simplicity and robustness. It minimizes jerk, the rate of change of force applied to the hand - the intuition being that efficient movements are smooth [6]. Let $\mathbf{z(t)} = [\mathbf{z_1(t)}\ \mathbf{z_2(t)}\ \mathbf{z_3(t)}]^{\mathrm{T}}$ be the hand's coordinates in 3D world coordinates, and t_s and t_e be the start and end times of the movement. It can be shown using Calculus of Variations that minimizing the jerk is equivalent to constraining z_1, z_2 and z_3 to be 5^{th} order polynomials in time t. Moreover, for starting and ending velocity and acceleration $= 0$, the loci are collinear. Let $\mathbf{y}(t) = [y_1(t), y_2(t)]$ be the projection of $\mathbf{z}(t)$ on the image plane. Of course, for collinear loci $\mathbf{z}$, the projected trajectory loci $\mathbf{y}$ are also collinear. Let $\dot{\mathbf{y}}(t)$ denote the projected velocity of the hand. For the i^{th} segment with t_s^i as the start time, we define $\dot{\mathbf{y}}(t) = B_i \vec{\tau}(t - t_s^i)$. B_i determines the dynamics of the i^{th} movement on the image plane; it remains constant for the movement. Assuming no mid-course corrections, the velocities within a segment are conditionally independent given B_i. Layer II in Figure 4 shows the equivalent Bayes net structure.

Layer III is the observation layer consisting of pose observations O_{p} and velocity observations O_{v}. Following the discussion in Figure 3, the dynamics of the hands are observed through low-level motion features. The Bayes net structure is shown in Layer III of Figure 4.

Summary of Recognition: Temporal segmentation is

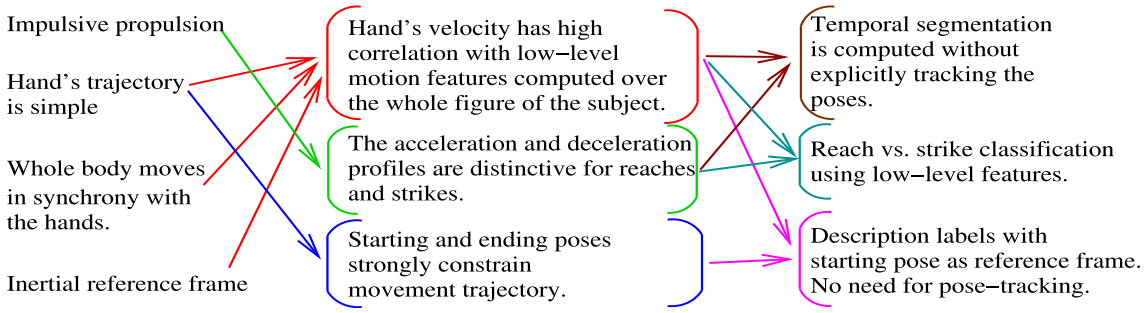


Figure 3. Observations in psycho-kinesiology have useful implications for visual analysis of ballistic movements. This is basis for the Bayesian model. Recognition is performed without tracking body-parts. Instead, low-level motion features are used, e.g., optical flow.

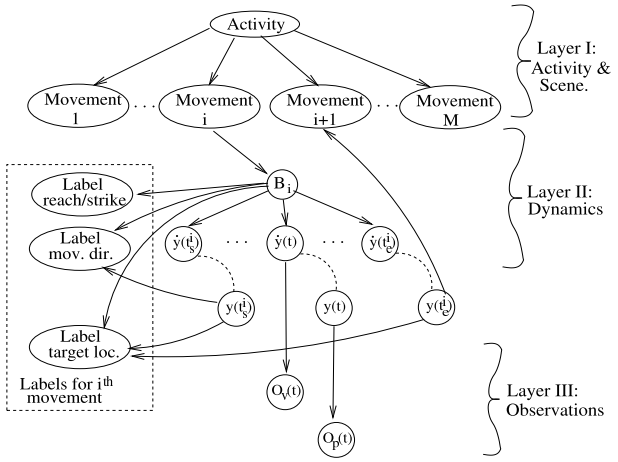


Figure 4. Bayes net for modelling ballistic movements. This is similar to the structure proposed by Bregler [3].

performed based on the constancy of B_i within segments. The Bayes net indicates that the state within a movement is determined by dynamics B_i, pose at the start of the movement, and the hand's target location; these constitute the ballistic verbs. Complex actions such as gestures are recognized by comparing sequences of B_i's.

2.1. Related Work

There have been a large number of studies on action recognition - see [8] for comprehensive surveys.

Bregler presented an approach for recognizing complex actions as a sequence of simpler atomic actions, called *movemes* [3]. Closely related are studies using Switching Linear Dynamical Systems (SLDSs) [12], and in general, body-part movement correlations.

Wilson and Bobick proposed Parametric Hidden Markov Models (P-HMMs) to handle variability in gestures [20]. P-HMMs would need a sufficient variety of training examples to generalize over all possible target locations. However, as they model the trajectory of movement, their approach may be used for recognizing different mass-spring movements, thus complementing our work.

Rao et al. proposed a scheme for segmenting human movement sequences based on the spatio-temporal curvatures of the hands' trajectories [11]. Weinland et al. segment continuous movement sequences using Motion History Volumes computed using 3D reconstruction[19]. The temporal segmentation in our approach uses single camera-view video and does not require tracking the hands.

State-of-the-art sub-space methods, e.g., [21, 15], have been developed to perform recognition robust to camera viewpoint and stylistic variation. Even for a stationary camera, two reach movements may have very different body-part trajectories if their target locations differ. Therefore, recognizing them involves generalizing over the dynamics in addition to the viewpoint. Our approach contributes to this aspect. A possible area of future study would to be to employ approaches such as [15] to explore the variation of matrix B_i w.r.t. subtle movement styles.

A study on ballistic movements in motion capture data [10] showed that the velocity magnitude profiles are distinctive for reach and strike actions. We extend it by addressing the more challenging problem of video-based analysis, and recognition of complex actions.

3. Features for Movement Dynamics

A. Trajectory-based Motion Features: These depend upon the spatial path followed by the hands during the movement. Examples include Space-Time Volumes (STVs) [21], Motion History Images (MHIs) [2, 19], Space-Time Gradients (STGs) [14]. They have been shown to be robust to noise, illumination variation, and small changes in view. We use Fourier-based MHIs and show that their effectiveness is enhanced by consideration of ballistic dynamics.

B. Velocity Magnitude Features: We propose a novel set of image motion features that isolate the manner of propulsion from the target of the movement. In brief, the velocity magnitude features are:

1. *Silhouette Deformation:* The subject's silhouette in each frame is computed using background subtraction followed by contour extraction [13]. A Distance transform $D_t(\mathbf{x})$ is computed on the image plane for the silhouette at each time instant t. The deformation of a silhouette at time t is measured by the histogram of the Chamfer distances of the points on the silhouette with respect to $D_{t-1}(.)$.

2. *Pixel-wise Frame Differences:* Let $\delta I_t(.)$ denote the thresholded image difference at time t. A distance map $D_t^{\delta}(\mathbf{x})$ is constructed from $\delta I_t(.)$. The pixel-wise frame difference feature is defined as the histogram of the distance

map values $\{D^{\delta}_{t-1}(\mathbf{x})|\delta I_t(\mathbf{x}) = 1\}$.

3. *Optical Flow:* A phase-based algorithm proposed by Gautama and Hulle [7] was used to compute optical flow. Let F_t denote the set of flow vectors at time t. We use magnitude of the net optical flow vector - $\| \sum_{\mathbf{f} \in F_t} \mathbf{f} \|$.

4. Temporal Segmentation

A continuous movement sequence is segmented such that the dynamics, B_i, within each subsequence is constant. Let the sequence be of time duration $[0, T]$. Let χ denote a partitioning of the sequence into n segments, $\chi = \langle \chi_0 = 0, \chi_1, \ldots, \chi_n = T \rangle$. The start of the i^{th} movement is $t^i_s = \chi_{i-1}$, and the end is $t^i_e = \chi_i$. The likelihood of the segmentation given the velocity observations, $p(\chi|O_{\text{v}})$ is modelled as $p(\chi|O_{\text{v}}) = p(B^*_1 \ldots B^*_n|O_{\text{v}})$, where B^*_i is the optimal dynamics for the i^{th} partition given the observations. By the conditional independence assumption

$$p(B_1 \ldots B_n|O_{\text{v}}) \propto \prod_i^n p(O_{\text{v}}(t^i_s) \ldots O_{\text{v}}(t^i_e)|B_i)\, p(B_i) \tag{1}$$

where $p(B_i)$ is the prior on the dynamics. The prior enforces constraints such as starting and ending velocity magnitudes should be close to 0. $p(O_{\text{v}}(t^i_s) \ldots O_{\text{v}}(t^i_e)|B_i)$ is the conditional probability of the velocity observations given the dynamics. We present an efficient algorithm for temporal segmentation based on the near straight line trajectories of ballistic dynamics, and the consistency of optical flow with the movement direction. Consider the i^{th} segment of duration $[t^i_s, t^i_e]$. Let the direction of movement of the hand be θ_i - this parameterizes the dynamics B_i. The likelihood of θ_i's fit to $O_{\text{v}}(t^i_s) \ldots O_{\text{v}}(t^i_e)$ is defined through potential functions on the weighted difference between the optical flow vectors and θ_i direction:

$$p(O_{\text{v}}(t^i_s) \ldots O_{\text{v}}(t^i_e)|B_i) = \prod_{t=t^i_s}^{t^i_e} \prod_{\mathbf{f} \in F_t} \exp - [\|\mathbf{f}\| - \mathbf{f} \cdot \hat{\mathbf{n}}(\theta_i)] \tag{2}$$

where $\hat{\mathbf{n}}(\theta) = \cos\theta \hat{i} + \sin\theta \hat{j}$. Maximizing with respect to θ_i, and plugging the optimal value of fit back in (1) gives

$$p(B_1 \ldots B_n|O_{\text{v}}) \propto \exp\left(\sum_{i=1}^{n} \left\| \sum_{t=t^i_s}^{t^i_e} \sum_{\mathbf{f} \in F_t} \mathbf{f} \right\| - \sum_{i=1}^{n} \sum_{t=t^i_s}^{t^i_e} \sum_{\mathbf{f} \in F_t} \|\mathbf{f}\| \right) \tag{3}$$

Notice that $\sum_{i=1}^{n} \sum_{t=t_s}^{t_e} \sum_{\mathbf{f} \in F_t} \|\mathbf{f}\|$ is a constant for the sequence, independent of the segmentation. Therefore, the optimality of the segmentation is given by

$$\sum_{i=1}^{n} \Psi(t^i_s, t^i_e) \quad \text{where} \quad \Psi(t^i_s, t^i_e) = \left\| \sum_{t=t^i_s}^{t^i_e} \sum_{\mathbf{f} \in F_t} \mathbf{f} \right\| \tag{4}$$

This is computed efficiently using Dynamic Programming (DP). In the implementation, we include the prior that the starting and ending velocity should be close to zero. It is valid to assume that ballistic movements have finite length, typically less than 3 seconds. Under this assumption, the DP algorithm can be computed online in realtime, and has $O(n)$ computational complexity. Figures 1, 8 and 7 show examples of obtained segmentations. Quantitative experimental results are given in Section 6.

Comparison with Velocity Magnitude based Movement Boundary Detection: Most ballistic movements, especially reaches and strikes, have zero velocity at the start and end of the segment [17, 6]. Therefore, it is natural to wonder if the temporal segmentation can be performed by simple movement begin-end detection. We trained a boosting based classifier [1] on velocity magnitude features to classify between segment boundaries and mid-flight points. Peaks in the classifier's output function should correspond to movement boundaries.

The results of the boosting-based movement begin-end detection were observed to be inferior to the DP-based segmentation algorithm. Reasons include the fact that the DP algorithm includes direction of movement and computes a globally optimal segmentation. Figure 5 illustrates this.

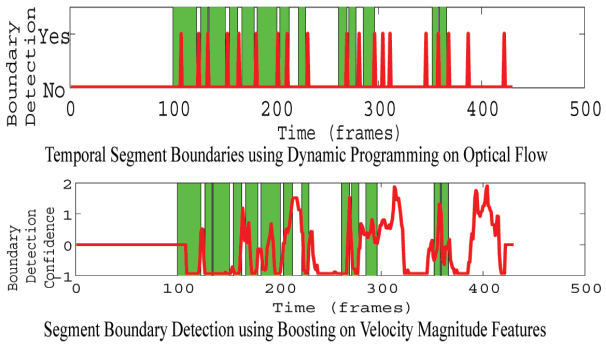


Figure 5. Segment boundary detection for a clip consisting of 13 reach movements. The time intervals of the movements are denoted with rectangles - gaps correspond to times of negligible motion. The boundaries of the temporal segmentation by the DP algorithm are marked with red spikes in top plot. They correspond well with the manually marked ballistic segment boundaries. The time-series plot for begin-end boosting detector's output (bottom red plot) is irregular, resulting in a number of missed and false boundary detections. The DP algorithm outperforms boosting.

5. Recognition

Annotation of Ballistic Verbs: Each movement's ballistic verb is a 3-tuple $\langle l_{\text{p}}, l_{\text{t}}, l_{\text{d}} \rangle$:

• l_{p} describes the manner of propulsion - *reach* or a *strike*.

• l_{t} describes the location of the target relative the person's starting pose. We use morphological labels: azimuthal location - *front, back, left, right* and *center*, and elevation - *ankle-level, knee-level, waist-level, chest-level* and *above-shoulder*.

• l_{d} describes the direction of movement relative to the person - *forward, backward, leftward, rightward, upward* and *downward*.

Propulsion: The velocity magnitude features are used to classify movements into reaches and strikes. As the dimensionality of features is large, we opted for a boosting-based

classifier [1]. The classification accuracy is high ($> 90\%$) and the method generalizes well over different target locations - Figure 6 illustrates typical output values of the classifier. Quantitative results are described in Section 6.

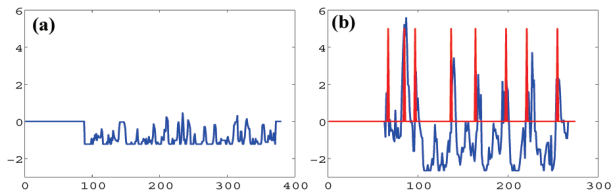


Figure 6. Strike detection using velocity magnitude boosting, classifier outputs as a function of time for: (a) clip with only reach movements, (b) clips with strike movements (hitting and throwing) manually marked with red spikes. The classifier output has peaks corresponding to ground truth strikes, and remains low elsewhere.

Target and Movement Direction: Let $\mathbf{y}_{\text{target}}$ denote the hand's target location in the image. If a ballistic segment with time-interval $[t_s, t_e]$, has been classified as a reach then $\mathbf{y}_{\text{target}} = \mathbf{y}(t_e)$. If the ballistic segment has been classified as a strike then $\mathbf{y}_{\text{target}} = \mathbf{y}(\text{time of highest strike confidence})$.

The subject's pose at the start of the movement, denoted by $O_{\text{p}}(t_s)$, provides spatial context to the target's location and the direction of movement. It is represented by the shape context of the person's silhouette and the head's gaze-direction.

Bayesian inference is used to compute the target and direction labels. Let $p_{\text{t}}(l_{\text{target}}|\mathbf{y}_{\text{target}}, O_{\text{p}}(t_s))$ denote the likelihood of label l_{t}, given $\mathbf{y}_{\text{target}}$ and $O_{\text{p}}(t_s)$. $p_{\text{t}}(.)$ is trained by collecting triplets of Shape-Contexts of the silhouettes at $\mathbf{y}_{\text{target}}$, the gaze-direction at the start of the movement, and the ground-truth labels for the target's locations.

The probability of l_{t} is computed by marginalizing over poses and target location estimates:

$$p(l_{\text{t}}) = \sum_{O_{\text{p}}(t_s)} \sum_{\mathbf{y}} p_{\text{t}}(l_{\text{t}}|\mathbf{y}, O_{\text{p}}(t_s)) p(\mathbf{y} \text{ is target } |O_{\text{p}}(t_s)) p(O_{\text{p}}(t_s)) \tag{5}$$

$p(O_{\text{p}}(t_s))$ is computed using the silhouette at the start of the ballistic segment and the confidences in the head's gaze-direction. $p(\mathbf{y} \text{ is target}|O_{\text{p}}(t_s))$ is estimated using skin detection and the motion features [13].

For computing direction labels, the target's location is replaced by B_i. The final label for each ballistic segment is computed as the maximum a posteriori probability estimate. Figures 1(a) 7 and 8 show some examples of computed labels. Quantitative results are described in the next section.

Sequences of Ballistic Movements: Complex actions are viewed as sequences of ballistic movements. We illustrate this with a gesture recognition application. Consider for example, the three gestures shown in Figure 1(d). The ballistic segmentation algorithm breaks them down into subunits, e.g., "raise right arm", "lower right arm". These are represented with MHIs [2, 19]. For simplicity, we use a Dynamic Time Warping algorithm to compare the sequences of ballistic segments. Results are described in next section. As part of future work, more sophisticated techniques such as HMMs may be explored.

6. Experimental Results

Annotation of Individual Ballistic Movements: We illustrate this on a database of reach and strike movements. 7 video sequences were collected depicting 67 reach instances performed by 6 subjects. A number of small objects such as pens, clips, etc. were placed on surfaces of varying heights in the scene. The subjects were asked to pick up and place the objects on random surfaces of their choice including the floor, in an area of 9×9 feet. Based on their own volition, the subjects stepped around, bent, used either of their hands, etc. They performed movements in rapid succession as well as with pauses. Movement instances in which the hands were occluded were ignored.

In a similar manner, we recorded 10 strike sequences depicting 68 instances of hitting and throwing performed by 4 subjects. The subjects were asked to strike and throw objects placed at heights varying from the ground to waist-level. Subjects punched, slammed down and slapped (forehand and backhand). The subjects struck and threw with all their might - one subject almost broke a garbage bin while slamming down on it!

The subjects consisted of 5 males and 1 female. The video resolution was 320×240, at 15 frames per second. The data-set is challenging as many movements are executed in rapid succession and at high speeds. The limbs are frequently inside the subject's silhouette, making pose-estimation difficult. There is significant motion blur during mid-flight. Table 1 shows the recognition results.

	Ground truth classes	
	Reaches	Strikes
1. Total number of instances (ground-truth)	67	68
2. Num. correctly segmented (percentage)	64 96%	68 100%
3. Num. classified as reaches (percentage)	60 **90%**	4 6%
4. Num. classified as strikes (percentage)	4 6%	64 **94%**
5. Correct reach/strike classifications & labelling of movement's direction and target location	56 **84%**	59 **87%**

Table 1. Recognition results for reaches and strikes

Segmentation results are shown in Row 2 of Table 1. Very few movements (3 of 135) were missed by the segmentation. The error in the boundary of the segments was in the range ± 3 frames (0.2 sec). A likely reason for this error is that the hand's velocity during the first few and last few frames of a movement segment is very low. Low level motion features are inadequate for such fine differentiation. Testing temporal segmentation through begin-end detection using velocity magnitude-boosting indicated it to be inferior to DP - 20% of reaches and 26% of strikes were missed. See Figure 5 for an example.

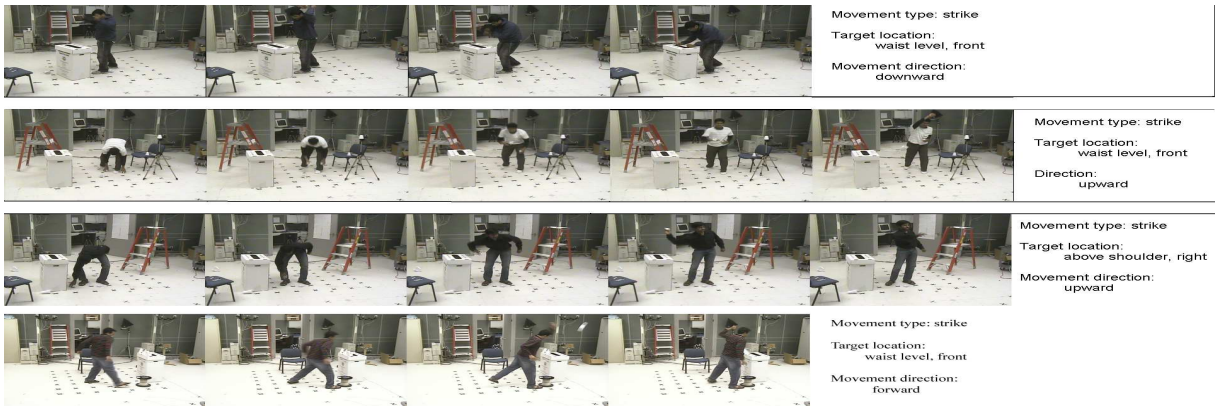


Figure 7. Labels generated for four strike movements. To save space, every third frame of the sequences are shown.

Figure 8. Labels generated for three reach movements. To save space, every third frame of the sequences are shown.

Reach vs. strike classification results are shown in Rows 3 and 4 of Table 1. The accuracy is high, the error rates being approximately 6%. In 2 of the cases in which strike movements were misclassified as reaches, the strike movement's duration was very small (2 to 3 frames). Due to the noise present in images and the subject's silhouette, it is difficult to reliably extract motion features for movements of such short duration. For comparison, a reach vs. strike boosting classifier was trained on MHIs. The obtained confidence scores have high overlap indicating unsuitability of MHI for the task, see Figure 10.

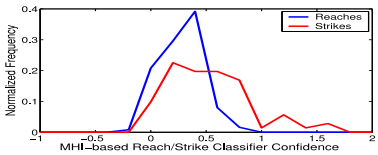


Figure 10. MHI based classifier response distributions for reaches (blue) vs. strikes (red). The high overlap indicates that MHI is inferior to proposed velocity magnitude features.

Target location & Movement direction results are shown in Row 5 of Table 1. The total number of reach movements that were correctly detected, classified and qualitatively labelled was 56 (84%). 2 of the target labelling errors were due to incorrect estimation of the hand's position at the end of the movement. The total number of strikes correctly detected, classified and qualitatively labelled was 59 (87%).

Gesture Recognition Results: The recognition of ballistic dynamics was tested on a gesture recognition application. We used the army signalling gesture dataset described in [16]. The dataset has 14 gestures, each performed 5 times by 5 subjects (350 sequences in total). Summary of the results using leave-a-subject-out protocol:

Method	Recognition rate
Nonparametric HMM [16]	84 %
MHI on full length of each gesture clip - no ballistic segmentation	73%
MHI on ballistic segments	**92%**

For the case of MHI on ballistic segments, recognition was performed by comparing test sequences with training sequences using DTW. Gesture labels were computed using Nearest Neighbor approach. Ballistic segmentation with MHI improves recognition rate over those reported in [16], as well as those obtained by solely MHI.

Example Videos:

(A.) *Tennis* video taken from [4] in Figure 1(b). The temporal segmentation algorithm computed consistent movement segments for the forehand swings; these correspond to poising, retracting hand, hitting ball, and follow through. All swings were detected corrrectly.

(B.) *Grapewine Pop* tutorial video in Figure 1(c). There is good correspondence between the ballistic segments computed for the four dance cycles. Notice that the movements are complex, multi-limbed and rapid; pose-tracking is likely to be hard. In the second missed boundary, the tutor almost jumped out of view, making optical flow erroneous.

(C.) *Furniture assembly* video in Figure 9. The video shows

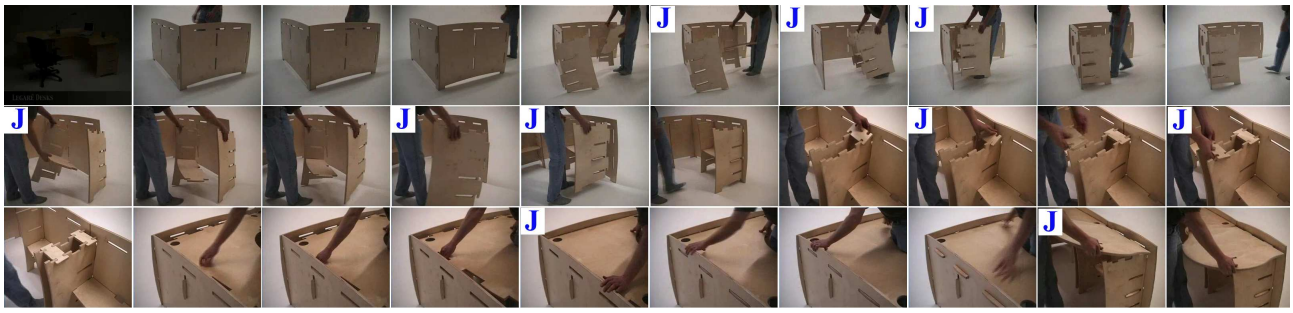


Figure 9. Ballistic segment boundaries for a furniture assembly video - courtesy www.legarefurniture.com, duration 1.5 min/2670 frames. 11 parts are joined to form a table. The video is automatically segmented into different join actions - frames at the time of join are marked with a 'J'. The first join is missed due to fade-in effect. Joins involving multiple movements are broken into multiple segments.

11 parts being joined to make a table. The segmentation detects 10 of these "join" actions - the first join is missed as it happens during a fade-in. Notice that there is severe occlusion of the pose due to camera view and the furniture parts.

INRIA XMAS Action Dataset: We also tested the approach on an action dataset presented in [19]. Fourier-based MHI features were employed for classification. Recognition improved from 83% to 87% upon including temporal segmentation by ballistic dynamics - indicating the utility of the concept. Example of ballistic segments for kicking

7. Summary

We presented a Bayesian framework for recognizing actions through ballistic dynamics. Comparative tests indicate that the approach is robust and effective. As an example, it enhances the performance of the popular MHI feature. Experiments with real-world videos highlight its consistent applicability.

Acknowledgements

The authors would like to acknowledge VACE for supporting the research. Yiannis Aloimonos, Ramani Duraiswami and David Jacobs provided many useful inputs during the project. Ashok Veeraraghavan, Pavan Turaga and Vinay Shet helped with the experiments.

References

[1] MSU Graphics and Media Lab, Computer Vision Group, http://graphics.cs.msu.ru.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–267, Mar. 2001.

[3] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR-1997*, 1997.

[4] A. E. Efros, A. C. Berg, G. P. Mori, and J. Malik. Recognizing action at a distance. In *ICCV'03*.

[5] M. Flanders, L. Daghestani, and A. Berthoz. Reaching beyond reach. *Exp. Brain Res.*, 126:19–30, 1999.

[6] T. Flash and N. Hogan. The coordination of arm movements: An experimentally confirmed mathematical model. *J. Neurosci.*, 5:1688–1703, Jul. 1985.

[7] T. Gautama and M. M. V. Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Trans. Neural Net.*, 13(5):1127–1136, 2002.

[8] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, Nov. 2006.

[9] P. Pigeon, S. B. Bortolami, P. Dizio, and J. R. Lackner. Coordinated turn and reach movements II: Planning in an external reference frame. *J. Neurophsio.*, 89:290–303, 2003.

[10] V. S. N. Prasad, V. Kellokompu, and L. S. Davis. Ballistic hand movements. In *Conf. Articulated Motion and Deformable Objects*, 2006.

[11] C. Rao, M. Shah, and T. Syeda-Mahmood. Invariance in motion analysis of videos. In *MULTIMEDIA '03: Proc. 11th ACM Int'l Conf. Multimedia*, pages 518–527, 2003.

[12] L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins, and J. M. Rehg. A data-driven approach to quantifying natural human motion. *ACM Trans. Graph.*, 24(3):1090–1097, 2005.

[13] A. Sepehri, Y. Yacoob, and L. S. Davis. Parametric hand tracking for recognition of virtual drawings. In *Proc. Fourth IEEE Int'l Conf. Computer Vision Systems 2006 (ICVS'06)*, page 6, 2006.

[14] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR-2005*, pages 405–412, 2005.

[15] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of human actions. In *ICCV-2005*, volume 1, pages 144–149, 2005.

[16] V. D. Shet, S. N. P. Vitaladevuni, A. Elgammal, Y. Yacoob, and L. S. Davis. Multi-cue exemplar-based nonparametric model for gesture recognition. In *ICVGIP'04*.

[17] I. Smyth and M. Wing, editors. *The Psychology of Human Movement*. Academic Press Inc., Orlando, FL 32887, 1984.

[18] Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectory in human multijoint arm movement. *Biol. Cybernetics*, pages 89–101, 1989.

[19] D. Weinland, R. Ronfard, and E. Boyer. Automatic discovery of action taxonomies from multiple views. In *CVPR-2006*, pages 1639–1645, 2006.

[20] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE PAMI*, 21(9):884–900, Sep. 1999.

[21] A. Yilmaz and M. Shah. Actions as objects: A nover action representation. In *CVPR-2005*, 2005.