Activity Recognition by Integrating the Physics of Motion with a Neuromorphic Model of Perception*

Ricky J. Sethi
UC Riverside
rickys@sethi.org

Amit K. Roy-Chowdhury
UC Riverside
amitrc@ee.ucr.edu

Saad Ali
Robotics Institute, Carnegie Mellon University
saada@cs.cmu.edu

Abstract

In this paper, we propose a computational framework for integrating the physics of motion with the neurobiological basis of perception in order to model and recognize human actions and object activities. The essence, or gist, of an action is intrinsically related to the motion of the scene's objects. We define the Hamiltonian Energy Signature (HES) and derive the S-Metric to yield a global representation of the motion of the scene's objects in order to capture the gist of the activity. The HES is a scalar time-series that represents the motion of an object over the course of an activity and the S-Metric is a distance metric which characterizes the global motion of the object, or the entire scene, with a single, scalar value. The neurobiological aspect of activity recognition is handled by casting our analysis within a framework inspired by Neuromorphic Computing (NMC), in which we integrate a Motion Energy model with a Form/Shape model. We employ different Form/Shape representations depending on the video resolution but use our HES and S-Metric for the Motion Energy approach in either case. As the core of our Integration mechanism, we utilize variants of the latest neurobiological models of feature integration and biased competition, which we implement within a Multiple Hypothesis Testing (MHT) framework. Experimental validation of the theory is provided on standard datasets capturing a variety of problem settings: single agent actions (KTH), multi-agent actions, and aerial sequences (VIVID).

1. Introduction

Understanding activities is intuitive for humans. From birth, we observe physical motion in the world around us and create perceptual models to make sense of it. Neurobiologically, we invent a framework within which we understand and interpret human activities [1]. Analogously, in this paper, we propose a computational model that seeks to understand human activity from its neural basis to its physical essence.

Motion underlies all activities; human activities, in fact, are defined by motion. The rigorous study of motion has been the cornerstone of physics for the last 450 years, over which physicists have unlocked a deep, underlying structure of motion. We employ ideas grounded firmly in fundamental physics that are true for the motion of the physical systems we consider in video.

Using this physics-based methodology, we compute **Hamiltonian Energy Signatures (HES)** for the various objects (either entire objects or the parts of a single object) involved in an activity, thus representing the motion of each object (or its parts) over the course of an activity as a scalar time-series. In addition, we develop a new distance metric, called the **S-Metric**, which also characterizes the global motion of the object, or the entire scene, with a single, scalar value (which can also be represented as a series of values if the total video is broken up into shorter time-segments since the S-Metric can be shown to be additive). Both the HES curves and the S-Metric provide a gist of the activity under consideration and offer a number of advantages for modeling actions and activities in videos.

In particular, we can show that the S-Metric is a proper distance measure over a metric space and we can also use basic physical principles to show that the S-Metric and HES are invariant under an affine transformation. This allows us to use the HES and S-Metric to *categorize* activities across different applications and domains (sparse/dense objects, high resolution, low resolution, etc.) in a moderately view-invariant manner without requiring separate heuristics (features or representations) for each. The HES and the S-Metric have distinct properties: the S-Metric can be used to characterize the *entire* scene with a single, scalar, global value; the HES time series, on the other hand, can characterize activities of individual objects.

Since the perception of activities involves the interpretation of motion by the brain, we embed the above physics-based motion models within a framework inspired by Neu-

*The authors at UCR were supported by NSF grant IIS-0712253, ARO grant W911NF-07-1-0485, and the DARPA VIRAT program.

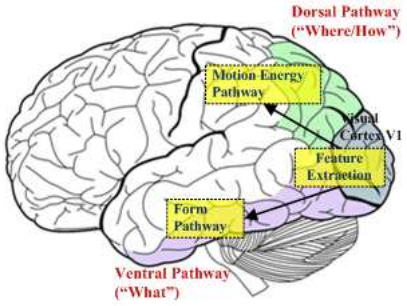

Figure 1. Feature extraction in V1 and then division along Motion Energy Pathway (Dorsal) and Form/Shape Pathway (Ventral)

robiology and **Neuromorphic Computing (NMC)**. The latest models for the perception and interpretation of motion by the brain are employed to present a novel technique for the representation and recognition of human actions. The neural basis for motion recognition, in fact, has garnered much attention of late.

Recent research, building upon the neurobiology of object recognition, suggests the brain uses the same, or at least similar, pathways for motion recognition as it does for object recognition [2, 3, 4, 5]. Visual processing in the brain, as shown in Figure 1, bifurcates into two streams at V1: a Dorsal Motion Energy Pathway and a Ventral Form/Shape Pathway [4, 6]. Although existing neurobiological models for motion recognition do posit the existence of a coupling or integration of these two pathways, they leave any specific mechanism for combination of the two pathways as an open question [2, 7]. This paper presents computational equivalents of these neurobiological models and applies them to real problems in computer vision.

Some researchers [8, 3] suggest the motion pathways integration is similar to object recognition; and since others [6, 9, 10] studying image-based recognition have been inspired by the success of biologically-motivated approaches for object recognition, we are similarly proposing the application of computational models that have proven effective for object recognition to motion recognition. In particular, neuromorphic computing [11, 12, 13] builds computational models for object recognition motivated by neurobiological pathways.

Building upon this and recent work in the neurobiological community which shows the dorsal and ventral processes could be integrated through a process of *feature integration* [14] or *biased competition* [15, 16, 17, 18, 19] as originally outlined by [20, 21], we propose a computational model for the fusion of the motion energy and form/shape pathways by representing this integration in a statistical **Multiple Hypothesis Testing (MHT)** framework.

Main Contributions: Building upon the fundamental principles of the physics of motion and the NMC model of perception, we present a novel framework for the modelling and recognition of actions and activities in video. Together, the HES curves and S-Metric give us an immediate sense of the gist of the motion energy of an activity since they are computed using global elements and features. For the Form/Shape element, we have the freedom to use different features (e.g., Histogram of Oriented Gradients for low-resolution video and shape/color for high-resolution video). Finally, we incorporate the Integration module by using a variant of the neurobiological ideas of feature integration and biased competition, which we implement using the MHT framework. This allows for hierarchical recognition, with gross recognition provided by the motion energy and a detailed analysis via the form information. Our representation provides flexibility since new approaches in low-level feature extraction can be employed easily within our framework. We present detailed validation of our proposed techniques and show results on querying a video database with complex activities.

2. Related Work

We build liberally upon theoretical thrusts from several different disciplines, including Analytical Hamiltonian Mechanics, Neuromorphic Computing and Neurobiology, and, of course, image analysis. The models developed for robotics in [11] provide the basic NMC architecture but are used more for image recognition and analysis. Similarly, Energy-Based Models (EBMs) [22] capture dependencies between variables for *image recognition* by associating a scalar "energy" to each configuration of the variables. Still others [23] take local and global optical flow approaches and compute confidence measures. Researchers have proposed computational frameworks for integration, e.g., [24], but they have also been restricted to the analysis of single images. The use of DDMCMC shown in [25], or its variants, might be an excellent Integration module application for future NMC-based research thrusts in situations where there is sufficient training data available.

In terms of human activity recognition [26], some of the cutting edge research uses the fusion of multiple features (e.g., [27]). Their approach to features fusion comes closest to the idea of combining features that express both the gist and the saliency, rather than multiple features where both express only one perspective. Our approach also draws inspiration from the method employed in [28], which detects global motion patterns by constructing super tracks using flow vectors for tracking high-density crowd flows in low-resolution. Our methodology, on the other hand, works in both high- and low-resolution and for densely- and sparsely-distributed objects since all it requires is the (x,y,t) tracks for the various objects.

In the Itti-Koch saliency model [13], the so-called saliency component (which corresponds to the form/shape pathway) is based on low-level visual features such as lumi-

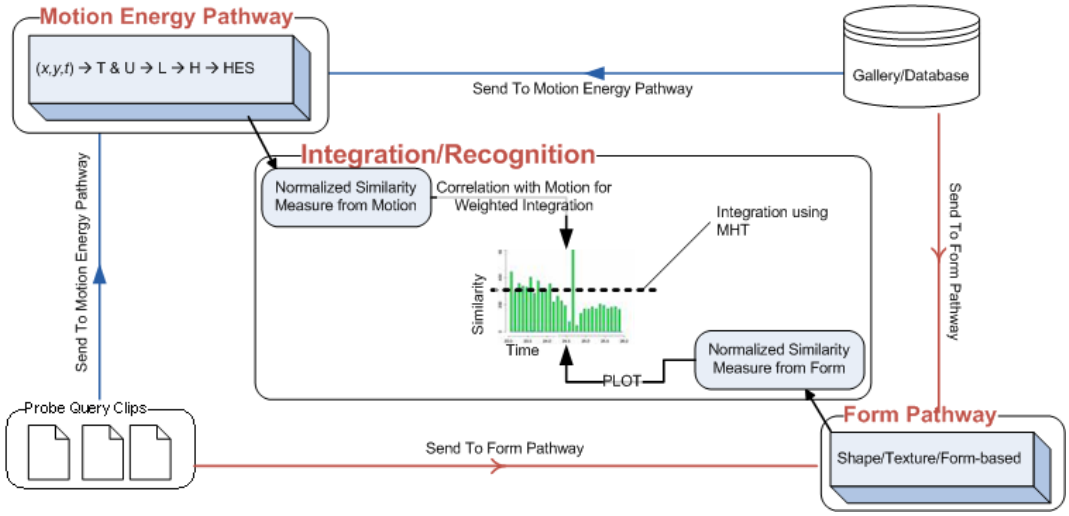

Figure 2. Proposed Framework for motion recognition by searching a database for a query: final recognition decision is made in the Integration module

nance contrast, color contrast, orientation, and motion with dyadic pyramids, while others use SIFT features, Kernel PCA, Harris corners, etc. [29]. The so-called gist component (which corresponds to the motion energy pathway) is usually computed as a low-level signature of the *entire* image with dyadic pyramids and Fourier energy [11] or color and texture or learnt statistical knowledge of the local features of the target and distracting clutter [29].

3. Proposed Framework for Activity Modeling

We propose an NMC-inspired approach for recognizing activities, using the global motion signature of the objects for the Motion Energy pathway (the Hamiltonian Energy Signatures and S-Metric, discussed below) and Local features for the Form component (the salient features, discussed below). The overall approach is shown diagrammatically in Figure 2.

3.1. Motion Energy Pathway: HES and S-Metric

One of the most fundamental ideas in theoretical physics is the *Principle of Stationary Action*, also known variously as Principle of Least Action as well as Hamilton's Variational Principle [30]. This is a variational principle that can be used to obtain the equations of motion of a system and is the very basis for most branches of physics, from Analytical Mechanics to Statistical Mechanics to Quantum Field Theory. We apply the idea of a function whose value remains constant along any path in the configuration space of the system (unless the final and initial points are varied) to Newtonian Mechanics to derive Lagrange's Equations, the equations of motion for the system being studied.

Following Hamilton's approach, we define **Hamilton's Action**, S, for motion along a worldline between two fixed physical events (not events in activity recognition) as:

$$S \equiv \int_{t_1}^{t_2} L(q(t), \dot{q}(t), t)dt \tag{1}$$

with q, the generalized coordinates [1], and L, in this case, the **Lagrangian** which, for a conservative system, is defined as:

$$L = T - U \tag{2}$$

where, T is the *Kinetic Energy* and U is the *Potential Energy*. The *Hamiltonian function*, derived from **Hamilton's Variational Principle**, is usually stated most compactly, in generalized coordinates, as [31]:

$$H(q, p, t) = \sum_i p_i \dot{q}_i - L(q, \dot{q}, t) \tag{3}$$

where H is the Hamiltonian, p is the generalized momentum, and $\dot{q}$ is the time derivative of the generalized coordinates, q. If the transformation between the Cartesian and generalized coordinates is time-independent, then the Hamiltonian function also represents the total mechanical energy of the system:

$$H(q(t), p(t)) = T(p(t)) + U(q(t)) \tag{4}$$

In general, we compute (3), which depends explicitly on time, but we can make the assumption (4) as a first approximation, in which the system can be idealized as a holonomic system, unless we deal with velocity-dependent or time-varying potentials.[2]

[1] Generalized coordinates are the configurational parameters of a system; the natural, minimal, complete set of parameters by which you can completely specify the configuration of the system.

[2] In fact, even when we cannot make those idealizations (e.g., viscous

Figure 3. Tracks to Hamiltonian to Phase Space: the phase space of a system consists of all possible values of the coordinates, which can be (*q,p*) or (*q,p,t*), for example; we may also look at modified phase plots of (*H,t*), (*H,q,p*), etc.

3.1.1 *Application to Activity Modeling*

Starting from these first principles, we develop a method to extract an abstract representation of the motion of the underlying physical systems we consider in video. The Hamiltonian in (3) is exactly what we utilize as the Hamiltonian Energy Signature (HES) for various objects (either entire objects or the parts of a single object) involved in an activity, thus representing the motion of each object over the course of the activity as a time series.

For example, if we track a person in video, we can compute these HES curves for the centroid of the person (considering the person as an entire object) or consider all the points on the contour of that person's silhouette, thus leading to a multi-dimensional time series (which can, for example, represent the gait of a person). Note that these HES curves can be computed in either the image plane, yielding the Image HES as used in this paper, or in the 3D world, giving the Physical HES, depending on the application domain and the nature of the tracks extracted. In either case, the Hamiltonian framework gives a highly abstract, compact representation for a system and can yield the energy of the system being considered under certain conditions.

We thus segment the video into systems and sub-systems (e.g., whole body of a person, or parts of the body) and, for each of those, get their tracks, from which we compute T and U, and use that to get the HES curve signature, which can then be evaluated further and the results analyzed accordingly. In the same manner, we compute the S-Metric from the tracks for the relevant time period by first computing the L from the T and U, as shown in Figure 3.

We end up with two quantities that provide a global description of the activity:

1. *HES* (3), which gives a simple, intuitive expression for an abstract, compact representation of the system; i.e., the characteristic time-series curves for each object
2. *S-Metric* (1), derived from L, which is the global, scalar signature of the system, possibly consisting of multiple objects

A system, in this sense, is defined according to the constraints of the video and the activity we are trying to identify. Thus, a system could be an object, like a car, represented as a particle, or a group of cars, where each car is a sub-system. Or, when the video permits, the system could be the car itself, with its various elements, the door, the hood, the trunk, being the sub-systems. Similarly, a human could be represented as a singular, particulate object that is part of a system of objects or as a system composed of sub-systems; i.e., when we can characterize their legs, arms, hands, fingers, etc. Thus, our approach is to segment the video into systems and sub-systems (e.g., whole body of a person, or parts of the body) and, for each of those, get their tracks, from which we compute T and U, and use that to get the HES curve signature, which can then be evaluated further in phase space [3] and the results analyzed accordingly, as shown in Figure 2.

Thus, we use the video to gain knowledge of the physics and use the physics to capture the essence of the system being observed via the HES and S-Metric. In order to compute the HES, we use the tracks from the video to compute the kinematic quantities that drop out of the Lagrangian formalism, thus giving a theoretical basis for examination of their energy from (*x,y,t*).

Examples: For example, in the general case when $U \neq 0$, the Lagrangian, T - U, of a single particle or object acting under a constant force, F (e.g., for a gravitational field, *g*, F=*mg*) over a distance, *x*, is:

$$\begin{aligned} &L(x(t), \dot{x}(t)) = \tfrac{1}{2}mv^2 - Fx \\ &\text{with } x = x_o + v_o t + \tfrac{1}{2}at^2 \text{ and } a = \tfrac{F}{m} \end{aligned} \tag{5}$$

We now use this Lagrangian to calculate Hamilton's Action for the general system:

$$\begin{aligned} &S = \int_{t_a}^{t_b} L dt = \int_{t_a}^{t_b} \left(\tfrac{1}{2}m(v_o^2 + 2v_o\tfrac{F}{m}t) - F\left(x_o + v_o t\right)\right) dt \\ &= \tfrac{1}{2}mv_o^2(t_b - t_a) - Fx_o(t_b - t_a) \end{aligned} \tag{6}$$

Using Hamilton's Variational Principle on (6) for a gravitational force yields (with *y* being the vertical position, which can be determined from the tracks):

$$H = T + U = \frac{1}{2}mv_o^2 + mgh = \frac{1}{2}mv_o^2 + mg(y_b - y_a) \tag{7}$$

Here, as a first approximation, we treat *m* as a scale factor and set it to unity; in future, we can estimate mass using the shape of the object or other heuristics, including estimating it as a Bayesian parameter. In addition, mass is not as significant when we consider the same class of objects.

For more complex interactions, we can even use any standard, conservative/non-conservative model for U (e.g., flows), we can define "generalized potentials" [31] and retain the standard Lagrangian, as in (2).

[3] The phase–space of a system consists of all possible values of the generalized coordinate variables q_i and the generalized momenta variables p_i. If the Hamiltonian is time-independent, then phase space is 2-dimensional, (*q,p*); if the Hamiltonian is time-dependent, then phase space is 3-dimensional, (*q,p,t*) [32].

as a spring with $U = \frac{1}{2}kx^2$ for elastic interactions, damped/driven harmonic oscillator for people following, etc.) whereas, for the simplest case of a free particle, we are left with just the T; this is equivalent to situations with videos of activities happening in far-field where we can only discern the motion of objects on a ground plane and we cannot imply any relationship between the objects. Even for these cases, we can also plot Hamilton's Action vs. time as the HES curve since the partial derivative of the Action is energy [30]. In addition, we can compute differences in the S-Metric between *multiple* objects by just subtracting their S Metrics, since Hamilton's Action can be shown to be additive.

Advantages: The approach we utilize to compute H and S is relatively straightforward, as shown in Figure 3: we find tracks for a scene, construct distance and velocity vectors from those tracks, and use these to compute HES vs. Time curves for each system or sub-system observed in the video. These HES curves and S-Metrics thereby yield a global signature, or gist, for the activity and allow us to characterize different activities or components of activities. For comparing the activities of different objects, we use the S-Metric/HES for each object. Since the HES is already a time-series, we can compare their characteristic HES curves using a Dynamic Time Warping (DTW) algorithm. We can also compare their full-fledged S-Metrics or, when we need a greater granularity of matches, we can segment the video into smaller time-intervals and compute the S-Metric piecewise for each of them, leading to a time-series in this case also. The additivity of S allows this and we can use DTW for matching the S-Metric sequences.

The main advantage of using the Hamiltonian formalism is that it provides a framework for theoretical extensions to more complex physical models. In addition, it can be shown that the Image HES allows us to recognize activities in a moderately view-invariant manner while the 3D Physical HES is completely view invariant; the invariance of both comes from the invariance of the HES to affine transformations. We show experimental validation of the invariance in Figure 4.

3.2. Form Pathway: Salient Features

Our construction provides flexibility on the form/shape pathway since new approaches in low-level feature extraction can be employed easily within our framework. For the present work, we use standard centroids, histogram of oriented gradients, and shape or color [33, 34]. In general, the Neurobiology literature suggests that orientation, shape, and color might serve as the best form/shape components [1, 7]. The only caveat is that the form component should maintain view-invariance so that the NMC model does not lose the view-invariance afforded by the HES/S-Metric during recognition.

Algorithm 1 Overview of the Multiple Hypothesis Testing (MHT) Algorithm.

Given a query clip, q, and a video DB, identify all occurrences of that query clip in the DB. Specifically, segment the video DB into c_i prospective clips whose Form Pathway normalized similarity measures compete for selection; winning values (those above the Tukey threshold) are then correlated with the Motion Energy Pathway normalized similarity measures.

1: Segment the Video DB into c_i prospective clips
2: Compute the plot of similarities for the Form pathway, $S_{Form}(q, c_i)$, and Motion Energy pathway, $S_{Motion}(q, c_i)$, by comparing each of the c_i prospective clips to the query clip, q
3: Compute Tukey (T) for the Form and Motion Energy distributions; this determines the Accept-Reject threshold for each
4: For each of the p Form similarity measures $\geq T_{Form}$, bias with the q Motion Energy similarity measures $\geq T_{Motion}$ (set all measures < T to 0)
5: Biasing is implemented as a pointwise correlation on the normalized similarity measures from Motion Energy and Form $\Rightarrow$ Survivors recognized

3.3. NMC Integration and Activity Recognition

The usual NMC tack is to integrate the form and motion energy pathways via the integration module, usually by weighting them. NMC and Neurobiologically-based approaches have examined different integration methodologies, including simple pointwise multiplication, as well as exploring more standard neurobiological integration mechanisms such as *feature integration* [14], in which simple visual features are analyzed pre-attentively and in parallel, and *biased competition* [20, 21], which "proposes that visual stimuli compete to be represented by cortical activity. Competition may occur at each stage along a cortical visual information processing pathway. The outcome of this competition is influenced not only by bottom-up, sensory-driven, activity but also by top-down, attention-dependent, biases".

We propose a computational approach to Integration that is a variant of these different methods and develop a computational framework that approximately mimics the neurobiological models of feature integration and biased competition [1, 15, 17]. There are a variety of different approaches that might be useful in simulating this integration. Since we deal with the case where training data is not available, we simulate this via a Multiple Hypothesis Testing (MHT) framework in which we first create "feature maps" [18] using the form/shape. A feature map is the distribution of form features with matching peaks using a similarity mea-

		Walking (A1)				Jogging (A2)				Running (A3)				Boxing (A4)				Clapping (A5)				Handwaving (A6)			
		V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
Walking (A1)	V1		0.003	0.002	0.001	0.067	0.112	0.133	0.063	0.595	0.537	0.984	0.976	0.750	0.546	0.884	0.852	1.000	0.980	0.935	0.833	0.457	0.316	0.395	0.382
	V2	0.003		0.004	0.004	0.095	0.121	0.142	0.073	0.763	0.691	0.820	0.816	0.780	0.600	0.898	0.870	1.000	0.983	0.943	0.853	0.521	0.396	0.466	0.454
	V3	0.002	0.004		0.001	0.061	0.088	0.123	0.043	0.513	0.455	0.803	0.790	0.756	0.559	0.887	0.855	1.000	0.981	0.936	0.837	0.471	0.335	0.411	0.399
	V4	0.001	0.004	0.001		0.061	0.100	0.125	0.050	0.549	0.492	0.951	0.941	0.760	0.566	0.889	0.858	1.000	0.981	0.937	0.840	0.480	0.347	0.422	0.410
Jogging (A2)	V1	0.067	0.095	0.061	0.061		0.010	0.014	0.004	0.413	0.400	0.544	0.613	0.707	0.472	0.864	0.826	1.000	0.977	0.924	0.804	0.371	0.221	0.310	0.299
	V2	0.112	0.121	0.088	0.100	0.010		0.001	0.015	0.354	0.327	0.430	0.526	0.698	0.459	0.860	0.821	1.000	0.976	0.921	0.798	0.357	0.209	0.299	0.283
	V3	0.133	0.142	0.123	0.125	0.014	0.001		0.021	0.340	0.313	0.413	0.522	0.694	0.454	0.857	0.818	1.000	0.976	0.920	0.795	0.352	0.209	0.295	0.277
	V4	0.063	0.073	0.043	0.050	0.004	0.015	0.021		0.409	0.378	0.579	0.577	0.711	0.480	0.866	0.829	1.000	0.977	0.925	0.807	0.380	0.231	0.318	0.307
Running (A3)	V1	0.595	0.763	0.513	0.549	0.413	0.354	0.340	0.409		0.028	0.021	0.074	0.664	0.412	0.843	0.800	1.000	0.973	0.911	0.774	0.307	0.173	0.241	0.224
	V2	0.537	0.691	0.455	0.492	0.400	0.327	0.313	0.378	0.028		0.006	0.024	0.670	0.418	0.846	0.804	1.000	0.974	0.913	0.778	0.312	0.170	0.249	0.231
	V3	0.984	0.820	0.803	0.951	0.544	0.430	0.413	0.579	0.021	0.006		0.018	0.649	0.393	0.835	0.789	1.000	0.971	0.907	0.763	0.294	0.177	0.222	0.209
	V4	0.976	0.816	0.790	0.941	0.613	0.526	0.522	0.577	0.074	0.024	0.018		0.655	0.400	0.839	0.794	1.000	0.972	0.909	0.768	0.294	0.162	0.225	0.209
Boxing (A4)	V1	0.833	1.000	0.846	0.852	0.590	0.541	0.508	0.609	0.390	0.424	0.333	0.366		0.014	0.002	0.001	0.008	0.007	0.004	0.000	0.040	0.141	0.070	0.078
	V2	0.785	1.000	0.810	0.822	0.508	0.458	0.427	0.530	0.310	0.339	0.258	0.286	0.012		0.033	0.029	0.051	0.048	0.040	0.025	0.000	0.082	0.020	0.027
	V3	0.854	1.000	0.863	0.867	0.628	0.580	0.546	0.646	0.432	0.466	0.373	0.408	0.003	0.027		0.000	0.001	0.001	0.000	0.000	0.055	0.163	0.087	0.097
	V4	0.849	1.000	0.860	0.864	0.620	0.572	0.538	0.638	0.423	0.457	0.364	0.399	0.002	0.025	0.000		0.002	0.002	0.001	0.000	0.052	0.158	0.083	0.093
Clapping (A5)	V1	0.867	1.000	0.875	0.877	0.653	0.606	0.572	0.670	0.460	0.495	0.401	0.437	0.007	0.035	0.001	0.002		0.000	0.000	0.003	0.066	0.175	0.099	0.109
	V2	0.865	1.000	0.873	0.875	0.649	0.602	0.568	0.667	0.456	0.490	0.396	0.433	0.006	0.034	0.001	0.002	0.000		0.000	0.002	0.065	0.173	0.097	0.107
	V3	0.860	1.000	0.868	0.871	0.639	0.592	0.558	0.657	0.445	0.479	0.386	0.422	0.004	0.031	0.000	0.001	0.000	0.000		0.001	0.060	0.169	0.093	0.103
	V4	0.846	1.000	0.857	0.862	0.615	0.566	0.532	0.633	0.417	0.451	0.359	0.393	0.001	0.024	0.000	0.000	0.003	0.002	0.001		0.051	0.156	0.081	0.090
Handwaving (A6)	V1	0.756	1.000	0.785	0.801	0.459	0.409	0.381	0.482	0.265	0.291	0.223	0.241	0.055	0.003	0.089	0.081	0.121	0.115	0.103	0.076		0.045	0.000	0.004
	V2	0.680	1.000	0.727	0.755	0.348	0.304	0.286	0.374	0.184	0.196	0.165	0.162	0.262	0.118	0.352	0.331	0.425	0.413	0.385	0.318	0.048		0.008	0.000
	V3	0.730	1.000	0.766	0.788	0.431	0.386	0.360	0.454	0.237	0.263	0.193	0.211	0.117	0.035	0.166	0.154	0.211	0.203	0.186	0.147	0.007	0.023		0.000
	V4	0.724	1.000	0.763	0.785	0.426	0.375	0.347	0.449	0.226	0.251	0.185	0.201	0.135	0.045	0.189	0.176	0.238	0.230	0.211	0.168	0.012	0.017	0.000	

Figure 4. KTH Distance Matrix where we highlight the lowest relative values in a row. This shows the matching of similar activities despite view changes with only a few exceptions to correct matching. Please note this is not necessarily symmetric because we do the analysis row-wise using training and classification.

sure between features. Then, these are fused via the multiple hypothesis testing framework.

This is accomplished as follows: suppose we are given a query clip that we want to match with a video database. We first divide the video into time segments and then compare each segment to the query clip. This similarity is in a suitable feature space; since the similarity measures are over a time window, they can be expressed as a distribution with a certain mean. So we need a method to see how significantly different means are. This is exactly what the Tukey test [35] does: it is a test that provides a way to determine whether a set of sample means are significantly different from each other.

The generalized MHT algorithm is shown in Algorithm 1. In our implementation, we plot the (normalized) similarity measures from the Form Pathway components for each time window. Then, for each of the peaks, or hypotheses, we test it with the Tukey. Finally, for those that pass the Accept-Reject threshold, T, we bias them by doing a pointwise correlation on normalized similarity measures between the mode of that sample distribution and the normalized Motion Energy Pathway gist value. This motion energy pathway value can be computed using DTW between the HES curve values of the query clip and potential match or the S-Metric difference between the query clip and the potential match. Finally, the results of this pointwise multiplication determine how many matches are recognized for the given query clip and video.

4. Experimental Results

We experimented with videos consisting of people, vehicles, and buildings, which encompasses a large class of possible activities. We used high-resolution and low-resolution video from standard datasets like KTH and VIVID. We also assumed tracking and basic object-detection to be available. We utilized these (x,y,t) tracks to compute the Kinetic (T) and Potential (U) energies of the objects (mass can be idealized to unity or computed from shape). The distance and velocity vectors derived from the tracks are thereby used to compute both the HES curves and the S-Metrics, which are then used as the Motion Energy Pathway of the NMC framework.

For the Form/Shape Pathway, we further used the tracks to get the histogram of oriented gradients [34] for the low-resolution case and utilized shape [33] in the high-resolution case. The histogram of similarities in each time window was computed. We then utilized Multiple Hypothesis Testing with the Tukey test to set the threshold for peaks in the distributions that might compete for selection/matching. We biased these peaks by doing pointwise multiplication with the motion energy gist computed earlier to make our final selections/matches.

4.1. Activity Modeling with HES/S-Metric

Here we first show an example of the characteristic HES curves and apply them to identify an exchange activity in video, as seen in Figure 5. In this scenario, two people, one of whom carries a box, approach each other. As they meet, they exchange the box and continue along their original

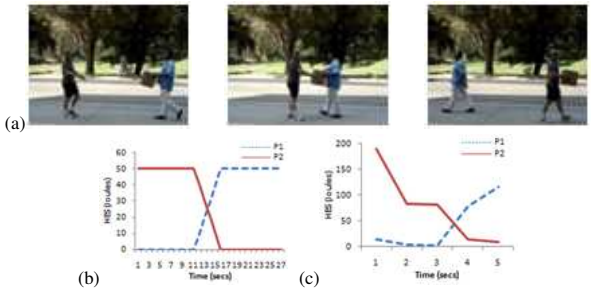

Figure 5. (a) Box Exchange experiment video: (b) Ideal vs (c) Actual Hamiltonian curves . Here we see two people exchanging a box in (a). Plots of the Hamiltonian equations of motion can give us a sense of the energies associated with this activity, both in the idealized case (b) and for the experimentally observed case (c).

S(1,2) = 0.05382	S(1,3) = 0.56237	S(2,3) = 0.63720

Table 1. S-Metric Distance between the three cars shows coupling between Car 1 and Car 2 (with a small distance) and the non-coupling between the others (showing larger distances).

paths unhindered. Some representative frames are shown in Figure 5.

Assuming an idealized exchange, a reasonable hypothesis might be to assume that the T and U of each person were identical, the only difference being the energies, both kinetic and potential, contributed by the box itself. In such a case, the energetics of this exchange can be represented as shown in the HES vs. Time plot in Figure 5, which shows both the idealized plot (a), where we assume identical people walking with identical speeds, as well as the actual, experimental plot (b).

The second example tracks three cars, where two cars maintain distance and one starts off together with them and then veers away, as shown in the frame in Figure 6. Since it involved more than two objects, we could then utilize the S-Metric to help characterize the gist of this system. For this experiment, we see the HES vs. Time curves for the two cars which follow all the way are highly correlated and the S-Metric (Table 1) calculated for them shows the coupling between Car 1 and Car 2 and the non-coupling between the others.

4.2. View Invariance of HES/S-Metric using KTH

The KTH dataset (`http://www.nada.kth.se/cvap/actions/`) contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. All sequences are taken over homogeneous backgrounds with a static camera with a 25fps frame rate. The sequences are downsampled to a spatial resolution of 160x120 pixels and have a length of four seconds on average. We use these to demonstrate the view invariance of the Motion Energy Pathway and present a distance matrix for all six actions in Figure 4.

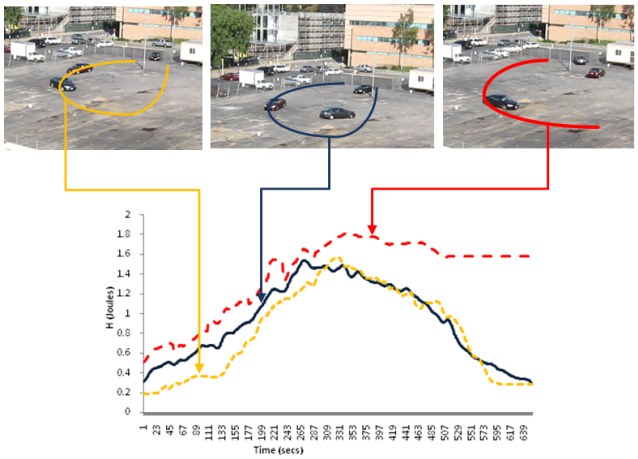

Figure 6. Two cars following; the first car, whose trajectory is labeled in orange, is the lead car and executes a U-turn; the second car, trajectory in blue, follows it and also makes a U-turn, whereas the third car, whose trajectory is in red, follows it for a while and then turns away.

As can be seen in Figure 4, there is significant matching between the same activity from different views, with lower scores indicating greater similarity. There are occasional exceptions as there are sometimes very few frames (as few as 10-20) with a sample rate of 25fps (i.e., there are too few frames for a completely reliable calculation of the HES curve) and the tracks can be tenuous at times. However, our model is able to distinguish between different activities, regardless of view, and matches the same activity, again, irrespective of the different view. We thus demonstrate the view invariance of the HES/S-Metric.

4.3. Querying using NMC on VIVID

We now show results on querying a large database using an example clip in order to provide evidence that integrating the Motion Energy gist (the HES/S-Metric) within the NMC framework performs better than just the Form/Shape alone. The database we use is the VIVID dataset (`http://www.darpa.mil/ipto/programs/vivid/vivid_approach.asp`) and we consider 6.5 minutes of it. The query clip was obtained from public domain data (example shown in Figure 6, above) and was 10 seconds in length. For want of space, we show results on a particular query, a car making a U-turn. There were 12 instances of it in the database. We show the Precision-Recall rates for this query for the Motion Energy Pathway, Form/Shape Pathway, and NMC-inspired Integration approaches in Table 2. As can be seen, the NMC-inspired approach increases the number of cases that can be retrieved.

	Precision	Recall
HES/S-Metric (Motion Energy)	0.62	0.63
HOG (Form)	0.67	0.55
NMC	0.754	0.75

Table 2. Precision, TP/FP, and Recall for Form, Motion Energy, and NMC

5. Conclusions and Future Work

The NMC-inspired framework and architecture we present provides a structured approach for a single, unifying framework for activity recognition that only requires tracks for the motion energy pathway; it is moderately view-invariant and can easily be generalized across different application domains and even applied to coupled systems, like cars chasing each other, exchanges, or interactions between sparse objects, and other systems without requiring separate heuristics for each. Our formulation takes an altogether novel approach whereby we attempt to create a theoretical framework inspired by the biological model and rooted in physics to gain insight into the problem of activity recognition in video. Future work will study how to obtain robust physics-based features, develop more complex physics-based models (e.g., using phase-space trajectories and Poisson Brackets), invert the two pathways' biasing in the Integration module, implement various other integration strategies, and use shape or learning algorithms to determine mass and potentials.

References

[1] E. Kandel, J. Schwartz, and T. Jessell. *Principles of Neural Science*. McGraw-Hill Medical, USA, 4th edition, 2000.

[2] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192, 2003.

[3] N. H. Goddard. *The Perception of Articulated Motion: Recognizing Moving Light Displays*. PhD thesis, University of Rochester, 1992.

[4] R. Sigala, T. Serre, T. Poggio, and M. Giese. Learning features of intermediate complexity for the recognition of biological motion. In *Artificial Neural Networks: Biological Inspirations, ICANN 2005*. Springer Berlin, 2005.

[5] M Riesenhuber and T Poggio. Hierarchical models for object recognition in cortex. *Nat Neuroscience*, 2:1019–1025, 1999.

[6] H Jhuang, T Serre, L Wolf, and T Poggio. A biologically inspired system for action recognition. ICCV, 2007.

[7] M. A. Giese. Neural model for biological movement recognition. In *Optic Flow and Beyond*, pages 443–470. Kluwer Academic Publihers, 2004.

[8] M. A. Giese. Neural model for the recognition of biological motion. In *Dynamische Perzeption 2*, pages 105–110. Infix Verlag, 2000.

[9] T Serre, L Wolf, and T Poggio. Object recognition with features inspired by visual cortex. CVPR, 2005.

[10] M Ranzato, F Huang, Y Boureau, and Y LeCun. Unsupervised learning of invariant feature hierarchies with application to object recognition. CVPR, 2007.

[11] R.J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. CVPR, 2007.

[12] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. pages 2049–2056. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[13] C. Siagian and L. Itti. Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. IROS, 2007.

[14] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cogn. Psychol.*, 12:97–136, 1980.

[15] G. Deco and E. Rolls. Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. *J Neurophysiol*, pages 295–313, 2005.

[16] D. M. Beck and S. Kastner. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, 2008.

[17] S Kastner and L Ungerleider. The neural basis of biased competition in human visual cortex. *Neuropsychologia*, pages 1263–1276, 2001.

[18] L. Yang and M. Jabri. Sparse visual models for biologically inspired sensorimotor control. pages 131–138. Proceedings Third International Workshop on Epigenetic Robotics, 2003.

[19] G. Deco and T. S. Lee. The role of early visual cortex in visual integration: a neural model of recurrent interaction. *European Journal of Neuroscience*, pages 1–12, 2004.

[20] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.*, 18:193–222, 1995.

[21] J.H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas v2 and v4. *J. Neurosci.*, 19:1736–1753, 1999.

[22] Y. LeCun, S. Chopra, M. A. Ranzato, and F. Huang. Energy-based models in document recognition and computer vision. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.

[23] A. Bruhn, J. Weickert, and C. Schnorr. Lucas/kanade meets horn/schunck: combining local and global optic flow methods. pages 211–231. Int. J. Comput. Vision, 2005.

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

[25] A. Yuille and D. Kersten. Vision as bayesian inference: analysis by synthesis. *TRENDS in Cognitive Sciences*, 10(7):301–308, 2006.

[26] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008.

[27] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *ICPR*. CVPR, 2008.

[28] M. Hu, S. Ali, and M. Shah. Detecting global motion patterns in complex videos. In *ICPR*, 2008.

[29] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. pages 1254 – 1259. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998.

[30] L D Landau and E M Lifshitz. *Course of Theoretical Physics: Mechanics*. Pergamon Press, Moscow,, 3rd edition edition, 1976.

[31] H. Goldstein. *Classical Mechanics*. Addison-Wesley, 1980.

[32] J. R. Taylor. *Classical Mechanics*. University Science Books, 2005.

[33] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with applications in human motion analysis. *PAMI*, pages 1896–1909, 2005.

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

[35] A Elliott. *Statistical Analysis Quick Reference Guidebook: With SPSS Examples*. Sage Publications, 2006.