Representing and Recognizing Human Motion: From Motion Templates to Movement Categories

(Tutorial Notes: *Digital Human Modeling Workshop*, IROS 2001)

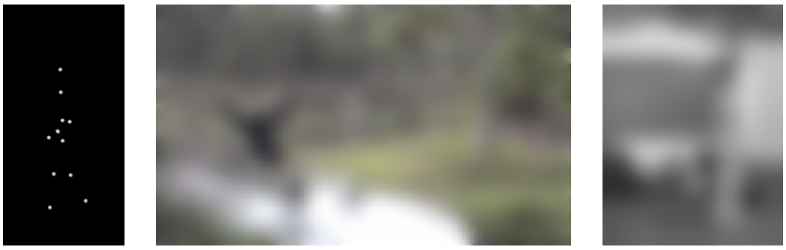
James W. Davis
Dept. of Computer and Information Science,
Center for Cognitive Science
583 Dreese Lab, 2015 Neil Avenue
Ohio State University
Columbus, OH 43210
`jwdavis@cis.ohio-state.edu`

1 Introduction

For reactive computers, embodied robots, or artificial creatures to effectively interact with us, it will be necessary for them to discern the various types of human movements that abound in the world. Additional applied significance for machines capable of human motion recognition can be found in automatic surveillance and monitoring systems, video content analysis, and perceptual user interfaces. The tracking and recognition of human motion, action, and events using computer vision has recently gained widespread interest in both academic and industrial research, with much emphasis on real-time systems (See [14, 56, 55, 34, 1, 68] for reviews of current research).

It is well known that certain types of biological movement patterns can be unambiguously recognized from their own organization of motion [46]. Though people generally cannot identify a *static* collection of bright dots in a dark scene as any meaningful object (See Fig. 1.a), they can easily interpret the *moving* dots as a set of lights attached to the joints of a walking person. People can further determine above chance the gender of the walker [17], even from viewing only two moving ankle points [47, 48]. In fact, babies have been found to stare longer at such biological point-light displays than at dots moving randomly [31]. This problem of recognizing biological movement has intrigued both psychologists and computer vision researches for decades, and has generated multiple controversial theories regarding reconstruction and recognition [36, 75, 46, 53, 70].

We have shown that even very low-resolution (blurred) video was sufficient for perceptual identification of common human (and animal) movements [22, 19]. A single frame from a blurred video sequence is shown in Fig. 1.b, which when subjects viewed this image (in color) reported seeing “an antelope in a field”, “an ice skater”, “maybe a plot of trees in the distance”, “the hood of a car”, or most originally “an impressionist painting viewed through a dirty window”. [What is your guess?] However, once the video is started from this frame, the immediate percept of a



(a) (b) (c)

Figure 1: (a) Video frame showing point-lights attached to a person's joints. (b) Initial frame of a blurred video sequence shown to viewers. Only after the blurred video is played does the action percept of a swinging gibbon become clear. (c) Blurred image from a sitting movement.

gibbon swinging along a rope is correctly identified. A similar non-identifiable image is shown in Fig. 1.c that can easily be recognized from the blurred *video* as a person sitting. As with the point-light displays, the movements throughout the sequence, rather than the static image features, were used as the dominant cue for recognition.

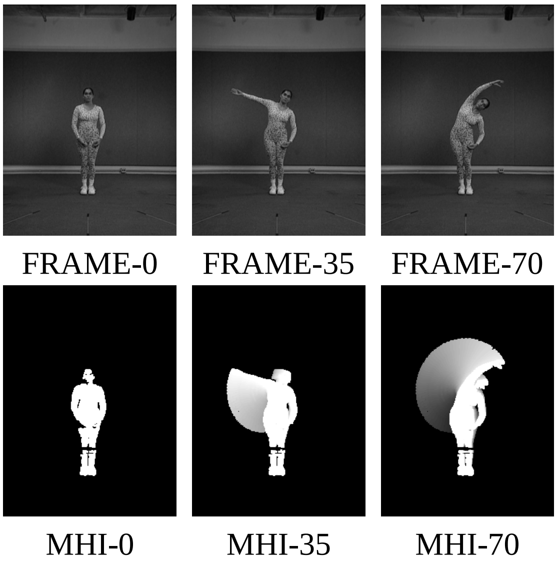
From these motivations, we developed two approaches to the representation and recognition of dynamic human movements. First we describe a real-time template-based method that compresses the holistic motion of an activity into a single motion history image. For recognition, higher-order moment features are computed from the template and statistically matched to trained models. In our second approach, we present a novel *categorical* framework for recognizing movements. Human movement is subject to a variety of physical and dynamic constraints, which together produce tight regularities in multi-dimensional feature spaces. Categories are fundamentally organized to exploit such correlations and thus these movement regularities offer a natural and descriptive basis for movement categorization.

2 Motion History Images

Rather than attempting the full three-dimensional reconstruction of the human form, we first developed a view-based approach to the representation and recognition of action that is designed to support the direct recognition of the motion itself (as motivated from the blurred video examples). The method focuses on accumulating and recognizing holistic "patterns of motion" rather than trajectories of structural features. The strength of the approach is the use of a compact, yet descriptive, real-time representation capturing a sequence of motions in a single static image (similar to [44]) called a *Motion History Image* (MHI) [22, 6]. The MHI is constructed by successively layering selected image regions over time (capturing the motion recency of pixels) using a simple update rule:

$$MHI_\delta(x,y) = \begin{cases} \tau & \text{if } \Psi(I(x,y)) \neq 0 \\ 0 & \text{else if } MHI_\delta(x,y) < \tau - \delta \end{cases} \qquad (1)$$

where each pixel (x,y) in the MHI is marked with a current timestamp τ if the function Ψ signals object presence (or motion) in the current video image $I(x,y)$; the remaining timestamps in the



FRAME-0 FRAME-35 FRAME-70

MHI-0 MHI-35 MHI-70

Figure 2: Top row: Keyframes of an arm stretching exercise movement. Bottom row: MHIs corresponding to keyframes in the top row.

MHI are removed if they are older than the decay value $\tau - \delta$. This update function is called for every new video frame analyzed in the sequence.

The function Ψ that selects a pixel location in the input image for inclusion into the MHI can be arbitrarily specified. Since the template representation captures both the position and temporal history of a moving object, many possibilities for selecting regions of interest are applicable. Detectors may include background subtraction, image differencing, optical flow, edges, stereo-depth silhouettes, flesh-colored regions, etc. With an *object* selection process for Ψ (e.g., background subtraction), the representation can accommodate slowly moving regions ($<$ 1 pixel/frame) that would otherwise be missed by image differencing or standard optical flow. For the results presented here, we used background subtraction and image differencing.

To illustrate the construction of an MHI, keyframes from a sequence of a person performing an "arm stretch" movement and the corresponding (cumulative) MHIs are presented in Fig. 2 (using background subtraction and $\delta = 2.33$ sec.). For display purposes the timestamp pixel values in the templates are linearly mapped to graylevel values 0–255. Here the brightness of a pixel corresponds to its recency in time (i.e., brighter pixels are the most current timestamps). Depending on the value chosen for the decay parameter δ, an MHI can encode a wide history of movement (See Fig. 3).

We also construct a binary cumulative motion image, referred to as a *Motion Energy Image* (MEI). The MEI indicates the *presence* of motion (the MHI describes the *recency* of motion), and is generated by thresholding the MHI above zero. Together, these images form temporal motion templates for representing human actions.

Similar use of templates for characterizing motion include work by [59, 51, 29], but are constrained to very particular domains (e.g., periodicity, facial motion). Our general template method is targeted at representing *arbitrary* human (and other) movements.

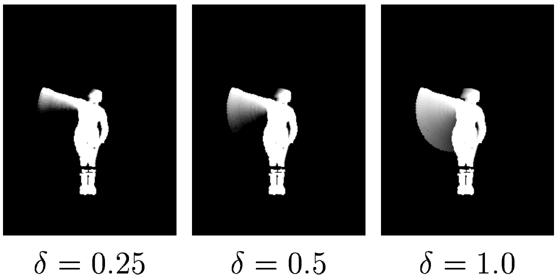


Figure 3: Effect of altering the decay parameter δ (in seconds) in Eqn. 1.

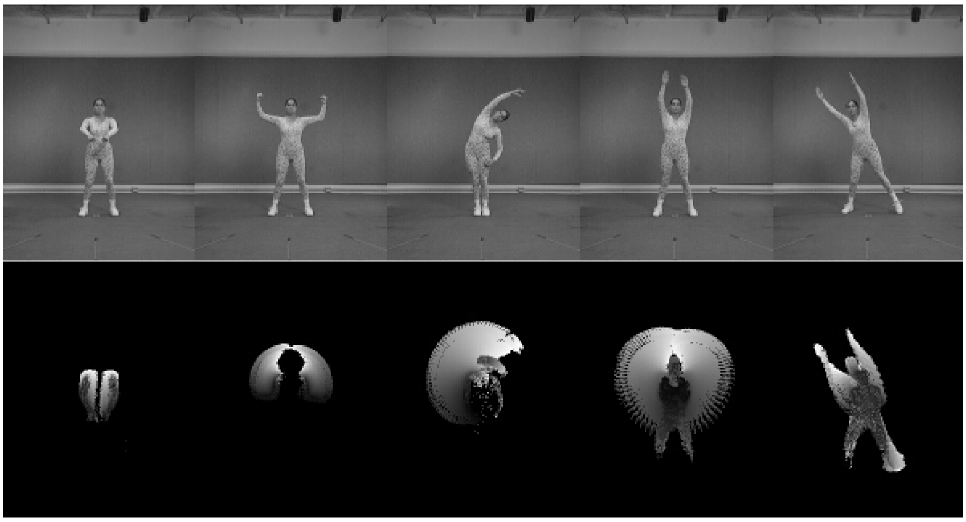


Figure 4: Examples from aerobics database and their MHIs (constructed using image-differencing).

2.1 Matching Motion Templates

To evaluate the power of the template representation, we recorded video sequences of 18 aerobics exercises performed by an experienced aerobics instructor. Seven views of the action ($+90°$ to $-90°$ in $30°$ increments in the horizontal plane) were recorded. For this experiment the temporal segmentation (duration of δ) was set manually, though an automatic method is described in [22]. For these experiments we used image-differencing for Ψ. A single key frame and corresponding MHI from the frontal view for 5 of the 18 aerobics exercises are shown in Fig. 4.

Given a set of MHIs and MEIs for each view/action combination, we compute statistical descriptions of the these images using 7 Hu moments [39] which are known to yield reasonable shape discrimination in a translation- and scale-invariant manner. To recognize an input action, the Mahalanobis distance [72] is calculated between the moment description of the input and each of the known actions.

2.1.1 Testing with One Camera

A new test subject performed each move and the input data was recorded by two cameras viewing the action at approximately 30° to left and 60° to the right of the subject. The MHI and MEI for each of the two views of the test input actions were constructed, and the associated moments computed.

Our first test uses only the left (30°) camera as input and matches against all 7 views of all 18 moves (126 total). We select as a metric a *pooled* independent Mahalanobis distance using a diagonal covariance matrix to accommodate variations in magnitude of the moments. Table 1.a displays the results. Indicated are the distance to the move closest to the input (as well as its index), the distance to the correct matching move, the median distance (to give a sense of scale), and the ranking of the correct move in terms of least distance.

The first result to note is that 12 of 18 moves are correctly identified using the single view. This performance is quite good considering the compactness of the representation (a total of 14 moments from two correlated motion images) and the large size of the target set. Second, the typical situation in which the best match is not the correct move, the difference in distances from the input to the closest move versus the correct move is small compared to the median distance. Examples of this include test moves 1, 9, 13, 16, 18. In fact for moves 1, 16, 18 the difference is negligible. Sometimes an alternative view of a different action projects into a template with similar statistics. For example, consider sitting and crouching actions when viewed from the front. The observed motions are almost identical, and the coarse template statistics do not distinguish them well.

2.1.2 Combining Multiple Views

A simple mechanism to increase the power of the method is to use more than one camera. Several approaches are possible. For this experiment, we use two cameras and find the minimum sum of Mahalanobis distances between the two input templates and two stored views of an action that have the correct angular difference between them, in this case 90°. The assumption embodied in this approach is that we know the approximate angular relationship between the cameras.

Table 1.b provides the same statistics as the first table, but now using two cameras. Notice that the classification now contains only 3 errors. The improvement of the result reflects the fact that for most pairs of this suite of actions, there is some view in which they look distinct. Because we have 90° between the two input views, the system can usually correctly identify most actions.

We mention that if the approximate calibration between cameras is not known (and is not to be estimated) one can still logically combine the information by requiring consistency in labeling. That is, we remove the inter-angle constraint, but do require that both views select the same action. The algorithm would be to select the move whose Mahalanobis sum is least, regardless of the angle between the target views. If available, angular order information — e.g., camera 1 is to the left of camera 2 — can be included. When this approach is applied to the aerobics data shown here we still get similar discrimination. This is not surprising because the input views are so distinct.

In response to the errors, the test subject for move 16 performed the move much less precisely than the original aerobics instructor. Because we were not using a Mahalanobis variance across subjects, the current experiment could not accommodate such variation. In addition, the test subject moved her body slowly while wearing low frequency clothing resulting in an MHI that has

Table 1.a: Single Camera Test					
	Closest	Closest	Correct	Median	Rank
	Dist	Move	Dist	Dist	
Test 1	1.43	4	1.44	2.55	2
2	3.14	2	3.14	12.00	1
3	3.08	3	3.08	8.39	1
4	0.47	4	0.47	2.11	1
5	6.84	5	6.84	19.24	1
6	0.32	10	0.61	0.64	7
Test 7	0.97	7	0.97	2.03	1
8	20.47	8	20.47	35.89	1
9	1.05	8	1.77	2.37	4
10	0.14	10	0.14	0.72	1
11	0.24	11	0.24	1.01	1
12	0.79	12	0.79	4.42	1
Test 13	0.13	6	0.25	0.51	3
14	4.01	14	4.01	7.98	1
15	0.34	15	0.34	1.84	1
16	1.03	15	1.04	1.59	2
17	0.65	17	0.65	2.18	1
18	0.48	10	0.51	0.94	4

Table 1.b: Two Camera Test					
	Closest	Closest	Correct	Median	Rank
	Dist	Move	Dist	Dist	
Test 1	2.13	1	2.13	6.51	1
2	12.92	2	12.92	19.58	1
3	7.17	3	7.17	18.92	1
4	1.07	4	1.07	7.91	1
5	16.42	5	16.42	32.73	1
6	0.88	6	0.88	3.25	1
Test 7	3.02	7	3.02	7.81	1
8	36.76	8	36.76	49.89	1
9	5.10	8	6.74	8.93	3
10	0.68	10	0.68	3.19	1
11	1.20	11	1.20	3.68	1
12	2.77	12	2.77	15.12	1
Test 13	0.57	13	0.57	2.17	1
14	6.07	14	6.07	16.86	1
15	2.28	15	2.28	8.69	1
16	1.86	15	2.35	6.72	2
17	2.67	8	3.24	7.10	3
18	1.18	18	1.18	4.39	1

Table 1: Results using (a) one camera and (b) two-cameras. Each row reports test move, distance to nearest move, distance to correct matching move, median distance, and correct move ranking.

large gaps in the body region. We attribute this type of failure to our simple motion analysis (image differencing); a more robust motion detection mechanism would reduce the number of such situations.

2.2 Motion Gradients

From Eqn. 1, the MHI layers the Ψ regions over time in such a way that the visual appearance of the layered regions gives the impression of motion directly from the intensity gradients in the template. It is quite apparent from MHI-70 in Fig. 2 that the upward progression of movement is captured in the dark-to-light intensity gradients. Since motion can be perceived from the displayed timestamp gradients in the template, one could convolve gradient masks with the timestamp values in the MHI to extract a motion vector at each pixel. This is similar in concept to computing normal flow along brightness contours [37]. We demonstrated this concept in [20, 18, 26].

Much like the work on hierarchical motion estimation, stereo matching, and image coding using image pyramids [4, 62, 13], we extend the original MHI representation into a pyramid to provide us with a means of addressing the gradient calculation of multiple image speeds using efficient, fixed-size gradient operators (e.g., Sobel masks). A pyramid permits the use of fixed-size gradient masks in each pyramid level (along with some constraints) to calculate motions of different speeds. The result is a hierarchy of motion fields where the resulting motion computed in each level is tuned to a particular speed (with faster speeds residing at higher levels). Due to the required anti-aliasing size reductions, the timestamped MHI cannot be transformed directly into a

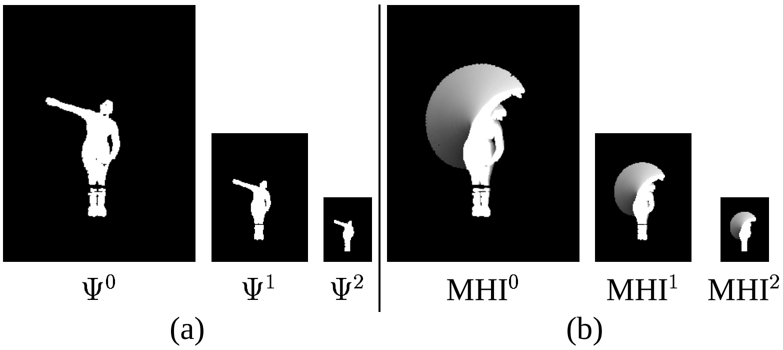


Figure 5: Image pyramids. (a) Thresholded Ψ pyramid of a person silhouette. (b) MHI pyramid created from (a).

pyramid. Instead, the Ψ image is first transformed into a pyramid and then each level from the Ψ pyramid is thresholded and used to update an MHI of that particular resolution (See Fig. 5).

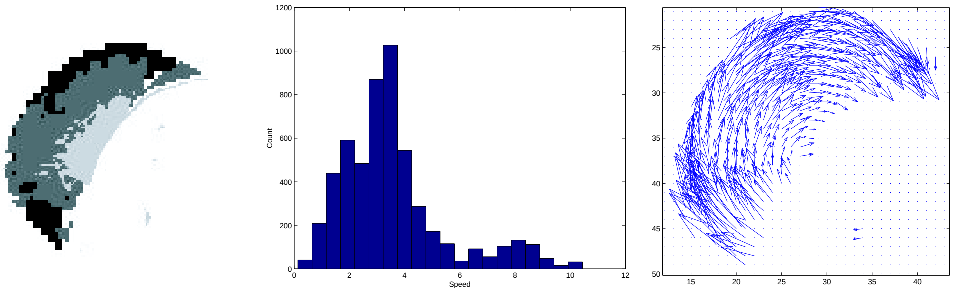
Before calculating the motion orientations and speeds in each level of the MHI pyramid using the gradient masks, a few constraints are initially required. First, the boundaries of the motion regions should not be examined by the gradient operators because of their adjacency to the null background values. We therefore impose an 8-connectedness test for each timestamp pixel to verify that it is "interior" to the motion region. Similarly, we do not examine the current timestamps (τ). We additionally constrain the F_x or F_y gradient to have an absolute value above some minimum to ensure that > 2 regions are being layered as a ramp (rather than a step of uniform region) within the gradient mask. For the implementation a threshold of 1/(2·FPS) is used.

Unlike the standard reintegration component (warp, expand, re-estimate) in the motion estimation of [4], we do not require an iterative propagation of the course-to-fine motion measures back to the size of the original MHI. Instead, for each pixel we 1) choose the finest resolution pyramid level that passes the gradient constraints, 2) compute the motion from that level using gradients, and 3) scale the result to the size of the original image. In Fig. 6.a we show the selected pyramid levels for each pixel in the arm raising MHI from Fig. 2. As expected with arc motion, the radially distant regions have a faster speed and are thus calculated at a higher pyramid level. In Fig. 6.b we plot a histogram of the computed speeds. The resulting motion field is displayed in Fig. 6.c which captures the overall expected pattern and organization of motions, with the larger motion vectors most radially distant from the point of rotation.

2.2.1 Motion Orientation Histograms

Once the motion field for the MHI has been computed (e.g., Fig. 6.c), several recognition methods could be applied to characterize particular movements. In the gesture recognition work of [33], a single histogram of image *edge orientations* of a user's hand was used to recognize various static gestures, with dynamic gestures formed by concatenating histograms of individual poses. We follow this approach and develop a *motion orientation* histogram by accumulating the orientations ϕ of the motion flow computed from the MHI pyramid (this method could be extended to incorporate speed as well).

We examined several different movements from the previous aerobic database of exercise ac-



(a) (b) (c)

Figure 6: (a) Pyramid levels assigned to pixels in the MHI (three-level pyramid with level-0=light-gray, level-1=medium-gray, and level-2=black). (b) Histogram of calculated pixel speeds. (c) Resulting motion field.

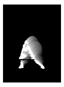
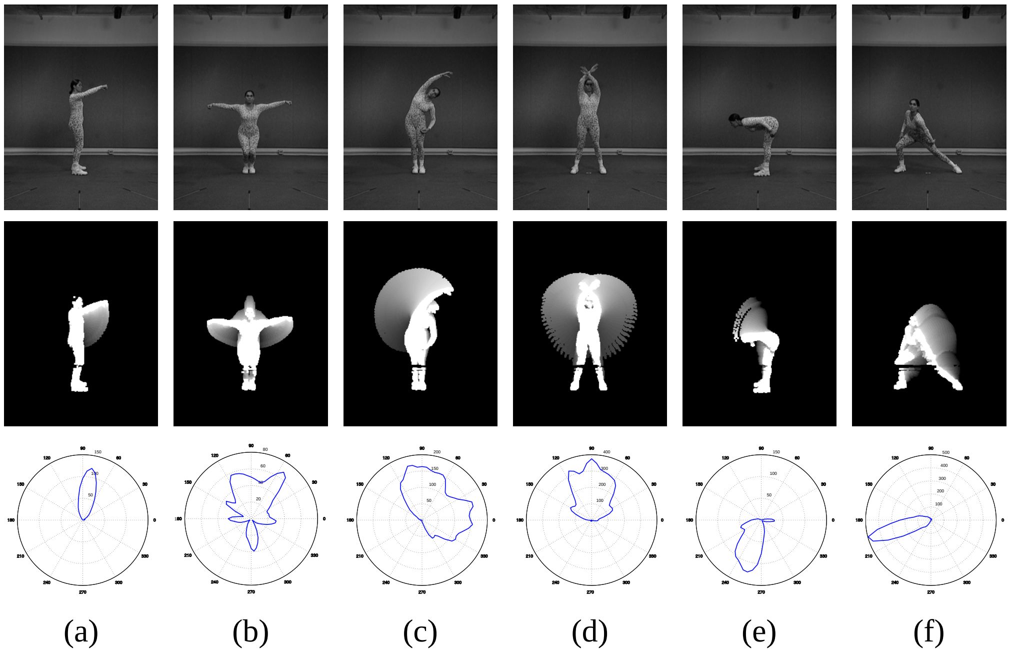
						
	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0.000	0.205	0.205	0.155	0.338	0.370
P_2	–	0.000	0.085	0.089	0.207	0.262
P_3	–	–	0.000	0.105	0.243	0.294
P_4	–	–	–	0.000	0.269	0.301
P_5	–	–	–	–	0.000	0.260
P_6	–	–	–	–	–	0.000

Table 2: Euclidean L_2 norm $||P_i - P_j||$ measures for normalized polar plots in Fig. 7.

tivities used in [22]. In Fig. 7, we present the final keyframe, MHI, and polar motion orientation histogram for those movements. The histograms are quantized by 5 degrees and smoothed with a 5-tap Gaussian filter. It is clear from the polar motion orientation histograms that these movements are significantly different from one another and that the histograms are useful for recognition. The comparative Euclidean L_2 norm distances between the movements in Fig. 7 are shown in Table 2.

2.3 Interactive Applications

The robustness of the moment-based MHI approach was demonstrated in two interactive systems. *The KidsRoom* [7, 8, 9] was an interactive fantasy-world playspace where children could participate in a reactive storybook narrative and interact with virtual characters (See Fig. 8.a). Our tracking system [43] and MHIs were used to monitor and recognize activities of the children. A newer version of this environment (*KidsRoom-2*) was constructed by Nearlife for a year-long exhibition in London's Millennium Dome. The other system was an interactive aerobics trainer [24, 76, 61, 52] that instructed and motivated people through an aerobic workout (See Fig. 8.b). The virtual instructor watched and recognized the exercise movements of the user employing our silhouette extraction technique [23] and MHIs. The output of the recognition system (e.g., jumping-jacks) and the responses of the virtual instructor (e.g., "*Good job!*") were coupled in a reactive system to



(a) (b) (c) (d) (e) (f)

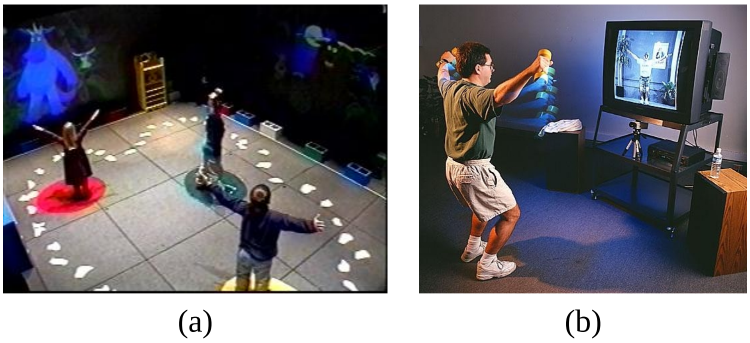
Figure 7: Top row: Key-frames from six aerobic exercise movements. Middle row: MHIs for the movements. Bottom row: Polar plots of motion orientation histograms computed using an MHI pyramid.

give the sensation that the instructor attended to the user.

The MHI motion template approaches are currently implemented in C++ using the highly optimized routines provided in the Intel Image Processing Library (IPL) and Open Source Computer Vision (OpenCV) library [42]. Many of the MHI functions and other necessary operators have already been incorporated into these packages [26, 11]. The main advantage to using the Intel libraries in terms of hardware is that faster processing is now accessible on standard PC-based platforms rather than on specialty systems or costly workstations. To date, there have been over 75K downloads of OpenCV. With this infrastructure, we are seeking stable real-time performance for a system that can be easily ported and made available to other researchers or developers.

2.4 Templates to Categories

The MHI framework was originally motivated by our perceptual ability to recognize particular actions using small amounts of structured motion information from limited visual input. For these percepts to be possible, we believe there must be correlated regularities in the motions to which our visual systems are tuned. Thus if we can identify such observable regularities that are characteristic to different *types* of human movement, we can encode these structures, as categories, directly into a computer vision system for the classification of our movements.



(a) (b)

Figure 8: (a) The KidsRoom. (b) Virtual aerobics trainer.

3 Movement Categorization

There has been much successful prior work in computer vision toward general *object* categorization. The related work on categorizing *motions* has mainly been concerned with low-level particle trajectory boundary characterizations (e.g., [67, 25]) or qualitative motion flow field descriptions (e.g., [5, 49]). We present here a novel categorization of *movement*. We refer to movement as the dynamic changing patterns of articulated limbs and bodies. We adopt movement categorization as a cognitively-motivated framework to provide insight into several computational issues related to constructing robust vision machines to recognize human movements. Our approach to movement categorization is based on the dynamic regularities of correlated movement features, rather than employing alternative rule-based or exemplar approaches to categorization [69].

3.1 Categories

The number of different movements in the world that must be dealt with by a vision system is perhaps infinite. Hence a basic task of any vision-based agent (biological or technological) having limited information processing capacities is to reduce the overwhelming number of movements to a more manageable number of *categories*[1]. Categorization is the process of treating certain non-identical items equivalently. Categories are organized to exploit correlations among properties, and thus enable computational and cognitive "economy" [66] for making decisions from partial sensory information. They provide a natural encoding of maximal information with the least amount of computational effort. Categorization has been shown to be a ubiquitous mechanism for perception in nature, from insects to primates to humans [38, 32, 73], and it is also a primary ontological component to Artificial Intelligence and knowledge engineering [71].

3.2 Categories as Regularities

The formation of perceptual categories requires that the world be structured with regularities available from visual input. Without such structure, it would be impossible for us to make any reliable inferences from our percepts. Given a configuration space Φ of dimension d parameterizable by R^d, a *regularity* $\rho \subset \Phi$ is a manifold of dimension $d' < d$ parameterizable by $R^{d'}$ [30]. Any

[1] A more general categorical statement is given in [12].

manifold is therefore a regularity with respect to some less-constrained configuration space. The smoothly varying structure of a regularity describes the dynamic “genericity” [30] of the category. Regularities are further characterized in terms of their predictive power, or modal nature [65, 45]. Specifically, this relates to how many additional property dimensions can be predicted from a single visual feature (quantitative prediction), and how much variance is inherent along the regularity (quality of the prediction).

Regularities of movement abound in our world. Their existence is mainly due to biomechanical constraints on body form, energy, and speed interacting with the laws of physics. A general movement regularity found across the animal kingdom is the dynamic Froude number $\frac{V^2}{GL}$, with running speed V, acceleration of gravity G, and leg length L. Froude values calculated for walking and running animals (and humans) can be used to show that different types of animals take relative strides and have similar gaits when their Froude numbers are the same [2]. Similarly, human locomotion is subject to a variety of physical and dynamic constraints [54, 41], which together produce tight regularities in multi-dimensional feature spaces. Only a few basic locomotion features, such as stride frequency, relative stride, and walking speed, are needed to produce a strong dynamic regularity (See Fig. 9). Interestingly, many movement regularities such as this can be represented with linear or log-linear parameterizations [54].

3.3 Movement Category Representation

We define a movement category as a parameterization of dynamic regularities together with any structural constraints:

Definition 1 *Movement Category = Dynamic Regularities + Structural Constraints*

The parameterization is used to generalize multiple exemplars, and can span several dimensions. The structural constraints allow for any fixed values to be included in the representation (e.g., people have *two* legs). This category description is similar to our previous structural and variable constraint model proposed for oscillatory motions [25, 19, 27].

To illustrate a basic movement category, consider the class of simple pendular motion of a swinging particle at the end of a light inextensible cord [64]. The *structural constraints* for this class of periodic movement include the choices for the particle mass M, the cord length L, and the gravity G. The *dynamic regularity* (for small amplitudes) consists of the correlation of the period of movement T to various ratios of L/G. This relationship can be parameterized by $T = \mathcal{F}(L, G) = 2\pi\sqrt{L/G}$. Assuming a gravitational constant of 9.8 m/s^2, this movement regularity can be represented in log-linear form: $\ln(T) = \frac{1}{2}\ln(L) + (\ln(2\pi) - \frac{1}{2}\ln(G))$. We can then predict (or verify) the period of movement for a pendulum from the observed length of the cord (mass does not effect the period). Our goal is to similarly represent various types of human movement as categories to provide a powerful and general mechanism for recognition.

3.4 Categories of Human Locomotion

We have selected the domain of human locomotion for a focused investigation of the movement category approach, though the framework remains general to other movement domains. After

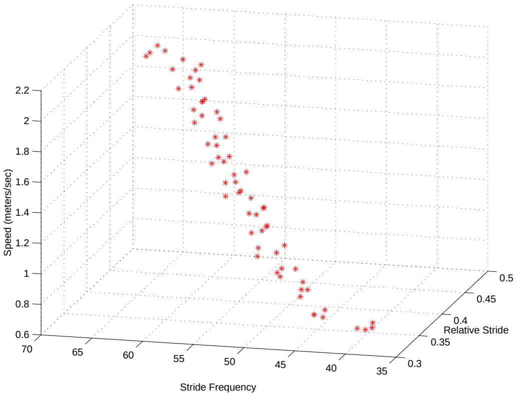


Figure 9: A dynamic locomotion regularity of stride frequency, relative stride, and walking speed for a single person.

a review of related work, we present preliminary results in the classification of typical walking movements and the sub-classification of children and adults.

Certain types of movement recognition can benefit from a model-based tracking pre-processor, but there are situations where direct movement recognition (with no part tracking) is applicable [22]. In [58, 59], the periodic motion of pixels (as created from walking) was analyzed throughout a sequence using Fourier techniques, followed by recognition of a feature vector derived from the motion in partitioned cells in the XYT volume. Another image-based periodicity approach is described in [16].

Another common approach to locomotion recognition includes the analysis of trajectory information derived from features of the walking body. In [74], the curvature trajectory of positions on a stick figure representation was examined using the Fourier transform to recognize cyclic walking movements. Principal component analysis and linear transformations are used in [77] to parameterize the temporal movement of body parts for the representation and recognition of walking. The frequency-based approach of [50] examined the phase relationships of periodic elements derived from optical flow for the task of person identification. Similarly, [40] recognized individual walkers by combining an Eigenspace transformation and a Fisher Linear Discriminate function on background-subtracted silhouettes. Also addressing identification, [57] used the spatio-temporal braided patterns of the legs within the XYT volume to detect walking, and then extracted features of the pattern to identify individuals. Lastly, [15] incorporated a sequence of body signature skeletons into an HMM framework to determine a posture transition path for recognition.

In contrast to the above recognition approaches, our interests focus more on the *categorization* of human movement to determine the *type* of locomotion (e.g., walking, running) with descriptive characterizations (e.g., age, gender, pace) that may even be further refined to allow person identification. We are motivated from the inherent biomechanical regularities associated with several low-level visual features, which form a grounded basis for our movement categorization. From the categories, we seek to determine several meaningful levels of descriptive analysis for the movement and person, rather than just a singular recognition match. This is a significant departure from the recognition work described above.

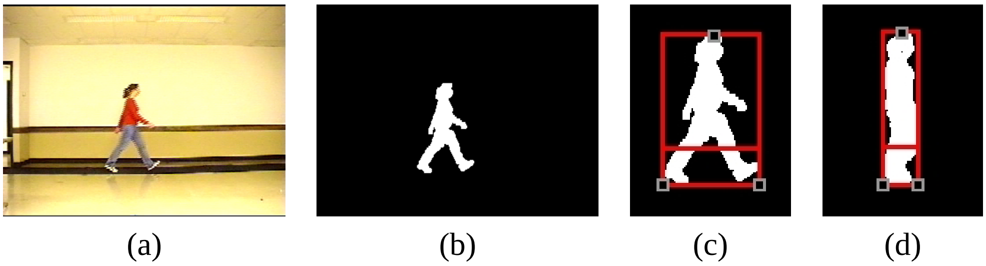


Figure 10: Silhouette generation and tracking. (a) Video frame of a walker. (b) Background-subtracted silhouette. (c) Maximal forward extension of lower quarter portion of silhouette. (d) Minimal forward extension. The movement features are extracted from the minimal/maximal extensions and can be computed using this method over several different viewpoints.

3.4.1 Categorization of Walking

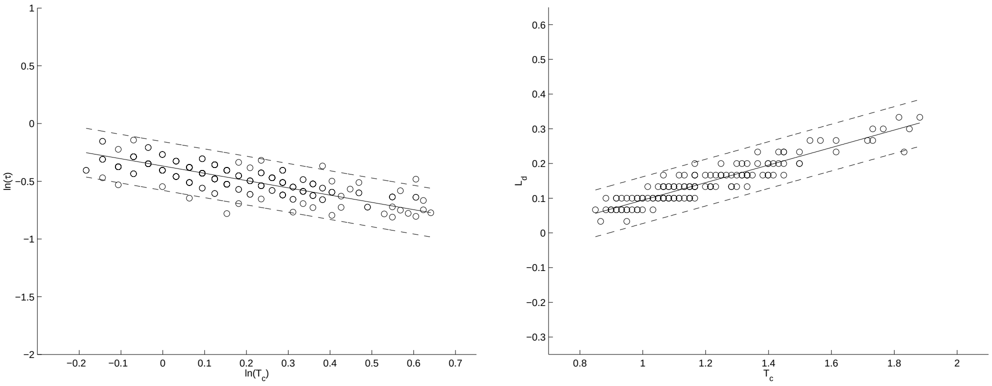
We can immediately determine an *atypical* walking pattern (e.g., limping) from inconsistencies of that movement with our strong perceptual category for *normal* human walking. Furthermore, the normal walking category itself contains several variations, all of which are considered to be typical. For example, though stride length increases with walking speed, the classification does not. We examine the smoothly changing visual behavior of locomotion over multiple walking speeds to capture the natural dynamics composing the movement regularities.

We relax the viewing constraint imposed on the recovery of spatial features (e.g., fronto-parallel alignment) by instead using salient temporal properties. We present three view-invariant temporal features for describing locomotion:

T_c = The cycle time of a leg.
τ = The ratio of stance-to-swing times of a leg.
L_d = The time difference from the onset of stance in leg 1 to the onset of swing in leg 2.

These features can be reliably obtained from low-level image analysis of people walking in video. We begin by extracting a silhouette of the person in each video frame with a background subtraction technique using RGB pixel differences, dilation, and removal of small pixel regions (See Fig. 10.a,b). We next form horizontal-motion trajectories using the leftmost and rightmost pixels within the bottom quarter section of the silhouette (See Fig. 10.c,d). The translation component is removed using the horizontal motion of the top-most detected head point of the silhouette. The movement features are directly calculated from the curvature extrema in these trajectories and are invariant to several changes in viewpoint (similar to the view-invariant trajectory representations in [63]). Using a synthetic walking sequence [60] rendered from multiple hip-level views ($0° - 90°$), this simple tracking approach was able to compute the same temporal events for a leg across views $45° - 90°$ (50% of the views).

To construct the locomotion regularities, we collected a video database of male and female walkers moving at slow, medium, and fast paces, and computed the movement features for each individual. Each walking cycle was analyzed separately with no averaging of features over multiple cycles. In Fig. 11.a, we present T_c vs. τ that characterizes the typical horizontal walking movements for a single leg. In Fig. 11.b, a second regularity of T_c (averaged T_c of the two legs)



(a) Single leg (b) Coupled legs

Figure 11: Movement regularities used for human locomotion categorization and classification. (a) Log-linear cycle time and stance-swing time ratio ($\rho = -.7674$). (b) Cycle time and leg delay ($\rho = .9269$). Computed linear prototypes with ±3σ are shown.

vs. L_d is shown. This regularity represents how the two legs move in relation to one another over multiple speeds.

To specify the degree of class membership of a new movement pattern to the category, we compute a perpendicular statistical distance of the new feature values to the regularity prototypes using

$$R_i = \frac{|\alpha_i X_i - Y_i + \beta_i|}{\sigma_i \sqrt{\alpha_i^2 + 1}} \tag{2}$$

with $\{X_1 = \ln(T_c), Y_1 = \ln(\tau), \alpha_1 = -.6299, \beta_1 = -.3680\}$ for the single leg regularity and $\{X_2 = T_c, Y_2 = L_d, \alpha_2 = .2524, \beta_2 = -.1580\}$ for the coupled legs regularity. The parameters for each regularity are determined from our training features using an Eigenvector line fitting process which minimizes the sum of squares of the *perpendicular* distances from the training points to the linear prototype [28]. The variance measures were computed as the overall variance along each regularity ($\sigma_1^2 = .0703, \sigma_2^2 = .0224$). We show the linear prototype models with $\pm 3\sigma$ class boundaries in Fig. 11. To be classified as typical walking, the regularity distances are thresholded and incorporated into the binary result M using the product classification rule [69] to discount movements having non-conforming properties:

$$M = \prod_i D_i, \qquad D_i = \begin{cases} 1 & R_i \leq 3\sigma_i \\ 0 & R_i > 3\sigma_i \end{cases} \tag{3}$$

We collected several new examples of typical and atypical walking sequences to test the approach. The set of typical walking included two people filmed a week apart from their training sequences, three new people not used in training, a 4-year-old child, a sequence each from the datasets of Little and Boyd [50] and Baumberg and Hogg [3], and a synthetic walking animation [60]. The atypical walking movements included a fast-paced walking video slowed down by 50%, a slow-paced walking sequence played at double the normal speed, a limping pantomime, a skipping movement, a marching walk, a light jog, a somersault, and a crawl. The classification results using Eqns. 2 and 3 were calculated for the sequences. Using the prior expected typical/atypical labels, 13 of 17 sequences were correctly classified. The exceptions included the skipping, marching walk, and light jog that were mistakenly determined to be within the range of normal walking. We attribute this behavior to the horizontal-only motion being categorized. Interestingly, the synthetic walking sequence was unexpectedly deemed to be atypical walking. Indeed, upon closer examination of the animation it did have a peculiar movement quality.

3.4.2 Distinguishing Children from Adults

We additionally posed the question of whether we could have a computer vision system further recognize the categorical differences between children and adults from the way they walk, rather than from static physical traits requiring a calibrated camera (e.g., body height).

Measurements of time and distance for each walking cycle represent the most basic descriptions that determine a particular gait [41]. A commonly used variable in describing locomotion is relative stride L', calculated by normalizing the person's stride length by stature (or leg length). The ratio is a dimensionless number, and thus can be used to compare the relative spatial configurations of children and adults in video. A descriptive temporal feature is stride frequency f (strides/min), computed from the inverse of the cycle time ($60/T_c$). To calculate these stride-based features, only the locations of the head and feet in each video frame are required, rather than the extraction of joint angles, limb lengths, or poses.

For this experiment, we collected video data of marked head and ankle positions from nine adults 30–52 years old and six children 3–5 years old. This particular age range for children was motivated by the reported biomechanical difference of children 3–5 years old as compared with adolescents and adults [35]. The adult subjects were recorded walking on a motorized treadmill at speeds ranging from 1.5–4.5 MPH, increasing in 0.2 MPH increments (individuals had different maximum speeds). The child movements were recorded as they walked back-and-forth across a room at different speeds. Treadmill and overground walking strides for an individual are not significantly different over the major range of our walking speeds [10].

The automatically computed ranges of stride frequencies of the children and adults were 55.3–89.8 strides/min and 36.7–73.3 strides/min, respectively. The relative strides for the children and adults were in the range 0.27–0.55. The natural log data for each cycle of each leg for all the walkers are depicted in Fig. 12. The general interpretations of this data are that the stride properties are positively correlated, and that when a child has the same (or larger) relative stride configuration as an adult, the child has a larger stride frequency. Since the data in Fig. 12 appears not fully Gaussian and almost linearly separable, we used a two-class (c1='Adult', c2='Child') linear perceptron classifier having the general form

$$d(\mathbf{x}) = \sum_{1}^{n} w_i x_i + w_{n+1} = 0, \quad \text{with} \quad \sum_{1}^{n} w_i x_i \underset{\text{c2}}{\overset{\text{c1}}{\gtrless}} -w_{n+1}. \tag{4}$$

The classification boundary computed for the database after 30K epochs using a Matlab Neural Network Toolbox implementation is shown in Fig. 12. When the entire dataset is classified using this discriminator, we receive 95% correct classification for the adults and 93% correct classification for the children. Three older children 5–6 years old were also tested and shown to have

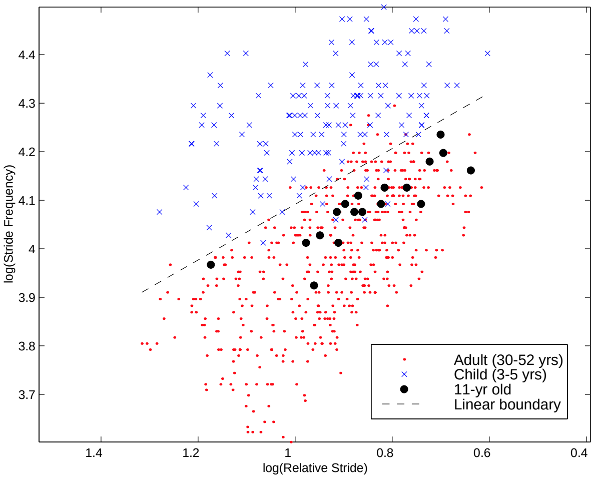


Figure 12: Relative stride length vs. stride frequency in children and adults (natural log values). A linear classification boundary is shown separating the two categories to a high degree. The walking movements of an 11-year old reside entirely in the adult category. Videos were recorded with a fronto-parallel viewpoint.

a larger percentage of movements than the younger children in the adult category (∼30%), suggesting the beginnings of a change in walking style. When an 11-year old was examined, his movements existed entirely within the adult category (See Fig. 12). Even though the two stride properties individually are not discriminatory, the *correlation* of the features is salient for classification. Given that only two movement features were used to characterize and differentiate the children from adults, the result is quite encouraging. Further details of this work can be found in [21].

4 Summary

We live in a dynamic world full of motion. Though multiple channels of visual information are utilized in understanding this world, motion itself is a powerful indicator of people and events. We described two approaches for the computer recognition of human activities.

We first presented a view-based approach to the representation and recognition of human movement. The basis of this real-time representation is a Motion History Image (MHI) that compresses the holistic motion of an activity into a single template. For recognition, higher-order moment features are computed from the templates and are statistically matched to trained models. The MHI was also transformed into an image pyramid to permit efficient fixed-size gradient masks to be convolved at all levels of the pyramid to extract motion information at a wide range of speeds. Polar motion orientation histograms were described as an additional means of recognition.

We next presented a novel categorical framework for representing and recognizing dynamic human movements. The approach is based on the inherent regularity in human movement to classify movement patterns to general movement categories. The approach was demonstrated with the classification of typical and atypical walking movements, and also by the discrimination of child and adult locomotion.

The task of constructing vision machines with even a fraction of our understanding for human (and animate) motions is indeed challenging. The outcome of this long-term research will have clear implications for computers, machines, and robots designed to monitor and interpret our actions.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: a review. In *Nonrigid and Articulated Motion Workshop*, pages 90–102. IEEE, 1997.

[2] R. Alexander. How dinosaurs ran. *Sci. Am.*, pages 130–136, April 1991.

[3] A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *Proc. Euro. Conf. Comp. Vis.*, pages 299–308, 1994.

[4] J. Bergen, P. Anadan, K. Hanna, and R. Hingorami. Hierarchical model-based motion estimation. In *Proc. Euro. Conf. Comp. Vis.*, pages 237–252, 1992.

[5] M. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. In *Proc. Comp. Vis. and Pattern Rec.*, pages 561–567, 1997.

[6] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 23(3):257–267, 2001.

[7] A. Bobick, J. Davis, and S. Intille. The KidsRoom: an example application using a deep perceptual interface. In *Proc. Wkshp. Percept. User Interfaces*, pages 1–4, 1997.

[8] A. Bobick, S. Intille, J. Davis, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. The KidsRoom: action recognition in an interactive story environment. *Presence: Teleoperators and Virtual Environments*, 8(4):367–391, 1999.

[9] A. Bobick, S. Intille, J. Davis, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. The KidsRoom. *C-ACM*, 43(3):60–61, 2000.

[10] W. Boda, W. Tapp, and T. Findley. Biomechanical comparison of treadmill and overground walking. In *Proc. Can. Soc. for Biomech.*, pages 88–89, 1994.

[11] G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. In *Proc. Wkshp. Applications of Comp. Vis.* IEEE, Dec. 2000.

[12] J. Bruner, J. Goodnow, and G. Austin. *A Study of Thinking*. Wiley, New York, 1956.

[13] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE transactions on communications*, com-31(4):532–540, April 1983.

[14] C. Cedras and M. Shah. Motion-based recognition: a survey. *Image and Vision Comp.*, 13(2):129–155, 1995.

[15] I. Chang and C. Huang. The model-based human body motion analysis system. *Image and Vision Comp.*, 18(14):1067–1083, 2000.

[16] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 22(8):781–796, 2000.

[17] J. Cutting, D. Proffitt, and L. Kozlowski. A biomechanical invariant for gait perception. *J. of Exp. Psych.*, 4(3):357–372, 1978.

[18] J. Davis. Recognizing movement using motion histograms. Technical Report 487, MIT Media Laboratory, Cambridge, MA, 1999.

[19] J. Davis. *Categorical organization and machine perception of oscillatory motion patterns*. Ph.D. Thesis, MIT Media Lab, Cambridge, MA, June 2000.

[20] J. Davis. Hierarchical motion templates for recognizing human motion. In *Proc. Wkshp. on Detection and Recognition of Events in Video*, pages 39–46. IEEE, 2001.

[21] J. Davis. Visual categorization of children and adult walking styles. In *Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication*, pages 295–300, 2001.

[22] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *Proc. Comp. Vis. and Pattern Rec.*, pages 928–934. IEEE, 1997.

[23] J. Davis and A. Bobick. A robust human-silhouette extraction technique for interactive virtual environments. In *Proc. Workshop on Modeling and Motion Capture Techniques for Virtual Environments*, pages 12–25. IFIP, 1998.

[24] J. Davis and A. Bobick. Virtual PAT: a virtual personal aerobics trainer. In *Proc. Wkshp. Percept. User Interfaces*, pages 13–18, 1998.

[25] J. Davis, A. Bobick, and W. Richards. Categorical representation and recognition of oscillatory motion patterns. In *Proc. Comp. Vis. and Pattern Rec.*, pages 628–635. IEEE, 2000.

[26] J. Davis and G. Bradski. Real-time motion template gradients using Intel CVLib. In *Proc. ICCV Workshop on Frame-rate Vision*. IEEE, 1999.

[27] J. Davis and W. Richards. Relating categories of intentional animal motions. Technical Report OSU-CISRC-11/00-TR25, Ohio State University, 2000.

[28] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[29] I. Essa and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. Int. Conf. Comp. Vis.*, pages 360–367, June 1995.

[30] J. Feldman. The structure of perceptual categories. *J. Math. Psych.*, 41:145–170, 1997.

[31] R. Fox and C. McDaniel. The perception of biological motion by human infants. *Science*, 218:486–487, 1982.

[32] D. Freedman, M. Riesenhuber, T. Poggio, and E. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001.

[33] B. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Proc. Int. Conf. on Auto. Face and Gesture Recognition*, 1995.

[34] D. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[35] D. Grieve and R. Gear. The relationship between length of stride, step frequency, time of swing and speed of walking for children and adults. *Ergonomics*, 5(9):379–399, 1966.

[36] D. Hoffman and B. Flinchbaugh. The interpretation of biological motion. *Biol. Cybern.*, 42:195–204, 1982.

[37] B. Horn. *Robot Vision*. MIT Press, Cambridge, 1986.

[38] G. Horridge and S. Zhang. Pattern vision in honeybees (*apis mellifera*): flower-like patterns with no predominant orientation. *J. Insect Physiol.*, 41:681–688, 1994.

[39] M. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, IT-8(2):179–187, 1962.

[40] P. Huang, C. Harris, and M. Nixon. Recognising humans by gait via parametric canonical space. *Artif. Intell. in Eng.*, 13:359–366, 1999.

[41] V. Inman, H. Ralston, and F. Todd. *Human Walking*. Williams & Wilkins, Baltimore, 1981.

[42] Intel Open Source Computer Vision Library,
`http://www.intel.com/research/mrl/research/opencv`.

[43] S. Intille, J. Davis, and A. Bobick. Real-time closed-world tracking. In *Proc. Comp. Vis. and Pattern Rec.*, pages 697–703. IEEE, 1997.

[44] R. Jain and H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 1(2):206–214, April 1979.

[45] A. Jepson and W. Richards. *Spatial Vision in Humans and Robots*, chapter What makes a good feature?, pages 89–125. Cambridge Univ. Press, 1991.

[46] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.

[47] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

[48] L. Kozlowski and J. Cutting. Recognizing the gender of walkers from point-lights mounted on ankles: some second thoughts. *Perception & Psychophysics*, 23(5):459, 1978.

[49] M. Langer and R. Mann. Dimensional analysis of image motion. In *Proc. Int. Conf. Comp. Vis.*, pages 155–162. IEEE, 2001.

[50] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre*, 1(2):2–32, 1998.

[51] F. Liu and R. Picard. Finding periodicity in space and time. In *Proc. Int. Conf. Comp. Vis.*, pages 376–383. IEEE, 1998.

[52] R. Martin. Virtual aerobics trainer shows future of computer UI. ABCNEWS.COM technology article, Nov. 6 1998.

[53] G. Mather, K. Radford, and S. West. Low-level visual processing of biological motion. *Proc. R. Soc. Lond. B*, 249:149–155, 1992.

[54] T. McMahon. *Muscles, Reflexes, and Locomotion*. Princeton Univ. Press, 1984.

[55] T. Moeslund. Computer vision-based human motion capture - a survey. Technical Report LIA 99-02, University of Aalborg, 1999.

[56] T. Moeslund. Summaries of 107 computer vision-based human motion capture papers. Technical Report LIA 99-01, University of Aalborg, 1999.

[57] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in XYT. In *Proc. Comp. Vis. and Pattern Rec.*, pages 469–474. IEEE, 1994.

[58] R. Polana and R. Nelson. Low level recognition of human motion. In *Workshop on Motion of Nonrigid and Articulated Objects*, pages 77–82. IEEE Computer Society, 1994.

[59] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *Int. J. Comp. Vis.*, 23(3):261–282, 1997.

[60] Poser. Curious Labs, Inc.

[61] M. Prigg. PC becomes home aerobics instructor. London Sunday Times innovation article, Nov. 15 1998.

[62] L. Quam. Hierarchical warp stereo. *Proc. Image Understanding Workshop*, pages 137–148, 1984.

[63] C. Rao and M. Shah. View-invariant representation and learning of human action. In *Proc. Wkshp. on Detection and Recognition of Events in Video*, pages 55–63. IEEE, 2001.

[64] R. Resnick, D. Halliday, and K. Krane. *Physics*. John Wiley & Sons, 1992.

[65] W. Richards, J. Feldman, and A. Jepson. From features to perceptual categories. In *Brit. Mach. Vis. Conf.*, pages 99–108, 1992.

[66] E. Rosch. *Cognition and Categorization*, chapter Principles of Categorization, pages 27–48. Lawrence Erlbaum Associates, 1978.

[67] J. Rubin. *Categories of visual motion*. Ph.D. Thesis, MIT Dept. of Psych., Cambridge, MA, Feb. 1986.

[68] M. Shah and R. Jain, editors. *Motion-Based Recognition*. Kluwer Academic, 1997.

[69] E. Smith and D. Medin. *Categories and Concepts*. Harvard University Press, Cambridge, MA, 1981.

[70] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *Proc. Comp. Vis. and Pattern Rec.*, pages 810–817. IEEE, 2000.

[71] M. Stefik. *Knowledge Systems*. Morgan Kaufmann, San Francisco, 1995.

[72] C. Therrien. *Decision Estimation and Classification*. John Wiley and Sons, Inc., New York, 1989.

[73] F. Tong, K. Nakayama, M. Moscovitch, O. Weinrib, and N. Kanwisher. Resonse properties of the human fusiform face area. *Cog. Neuropsych.*, 17(1/2/3):257–279, 2000.

[74] P. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection for motion based recognition. *Pat. Rec.*, 27(12):1591–1603, 1994.

[75] J. Webb and J. Aggarwal. Structure from motion of rigid and jointed objects. *Artif. Intell.*, 19:107–131, 1982.

[76] Fast forward: virtual workouts. WHDH-TV Channel 7 Boston news feature interview, Dec. 29 1999. First aired June 8, 1998.

[77] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 73(2):232–247, 1999.