OSU DEPT. OF COMPUTER AND INFORMATION SCIENCE TECHNICAL REPORT OSU-CISRC-1/03-TR2.

A RELIABLE-INFERENCE FRAMEWORK FOR POSE-BASED RECOGNITION OF HUMAN ACTIONS

James W. Davis *Ambrish Tyagi*

Dept. of Computer and Information Science
Ohio State University
Columbus, OH 43210 USA

ABSTRACT

We evaluate the feasibility of action recognition from individual poses. Our approach is formulated in a probabilistic reliable-inference framework using *a posteriori* class ratios to verify the saliency of a pose before committing to any action classification. The framework is evaluated in the context of walking, running, and standing poses at multiple views and compared to ML and MAP approaches. Results indicate that these actions can be reliably discriminated from a single image.

1. INTRODUCTION

Advanced video surveillance systems will require the capabilities of *detecting* the presence of people, *tracking* their movements, and *recognizing* their behaviors and actions. With respect to action recognition, how long of a video sequence is needed for reliable computer detection of basic human actions such as walking and running? Typically, analysis over several frames is employed to construct representations for recognition (e.g., matching trajectories or detecting characteristic motion patterns [1]). But how reliable could a system perform when limited to analysis of only *one* video frame? In other words, how reliable could individual poses be toward computer recognition of actions?

Clearly, recognition of basic activities from only a single frame (or few frames) of video would be advantageous to automatic video-based surveillance systems, especially for systems with limited computational processing time scheduled per camera. Also consider small unmanned aerial vehicles (UAVs) equipped with cameras. The UAV's view area is constantly changing, and therefore immediate decisions about the activity in the scene are desirable. Even when longer duration video is available, rapid action detection may be particularly helpful in bootstrapping more sophisticated action-specific tracking or recognition approaches.

We present an approach to pose-based recognition of actions that is formulated in a probabilistic inference framework that first verifies the reliability of a pose using an *a posteriori* comparison of the target classes before committing to any particular action classification. If the pose is deemed unreliable (too confusing), then no action classification takes place. Other probabilistic methods such as *maximum likelihood* (ML) and *maximum a posteriori* (MAP) instead perform a forced-choice classification regardless of the saliency of the input.

The advantage of the proposed method is that the system only makes classifications when it believes the pose is "good enough" for discrimination between the possible actions. This is particularly favorable when there is a high cost for making errors and low (or no) cost for passively waiting for another pose image to arrive (advantageous with real-time video).

We evaluate the reliable-inference framework using the task of discriminating walking, running, and standing poses from multiple viewpoints. We present results examining the framework and make comparisons to alternative ML and MAP approaches. We also examine the discrimination ability as a function of viewpoint to determine the best camera locations. We show that low Bayes error rates can be achieved for recognizing walking, running, and standing from single poses. To further illustrate the detection and elimination of confusing poses, we present results discriminating walking pace (slow, medium, fast).

We begin with a review of related pose-based detection and recognition methods (Sect. 2). Next we present the reliable-inference framework (Sect. 3), including methods for probabilistic modeling of the classes and for recognition. We then describe how the action database was collected and what features were chosen to represent the poses (Sect. 4). The experimental evaluations are presented (Sect. 5), followed by a summary and conclusion (Sect. 6).

2. RELATED WORK

In [2], wavelets were used to learn a characteristic pedestrian template for detecting people in cluttered scenes. The training set consisted of front- and rear-view color images of people in natural scenes (images were clipped and scaled

to a fixed size). The system was trained with additional negative examples using bootstrapping, and support vector machines were employed for classification using the wavelet coefficients as features.

A hierarchical coarse-to-fine template approach with radial basis function (RBF) classification was used in [3] to also detect pedestrians. The template hierarchy was constructed automatically from examples using refinement clustering of images into prototypes (using the Chamfer distance). During matching, a distance threshold between prototype candidates and the new image were used to prune the search through the hierarchy. Candidate matches were then verified using the RBF classifier.

For discriminating humans and vehicles, two simple properties (dispersedness, area) were used by [4] to classify regions selected from image differencing. To aid in temporal consistency of the labeling, a classification histogram was computed to accumulate over time the class labels assigned to a particular region. If the target region persisted for a given duration, the peak in the classification histogram was used to label the object.

A point distribution model was used in [5] to model the changing silhouette contour shape of a walking person (at different views) with cubic B-splines. Principal components analysis (PCA) was used to capture the significant modes of variation in the feature vectors for the various contour shapes. The direction of walking for each pose was appended to the feature vector to enable the estimation of the walking direction for new silhouette poses after reconstruction from the PCA space.

In [6], 2-D pose estimation from image silhouettes was cast in a general unsupervised learning framework using EM-based clustering to build a mapping between low-level moment features and 2-D joint positions. The model was trained using synthetic silhouettes rendered from multiple viewpoints and was demonstrated with pose recovery on both artificial and real images.

Our approach similarly employs an EM-based clustering of silhouette poses using moment features as in [6], but unlike the above approaches, we formulate the classification task as a probabilistic decision employing reliable-inference to classify only the most discriminating poses. Our method is designed to ignore unreliable poses during immediate decision-making, rather than requiring temporal consistency before classification (as in [4]). We also examine multiple viewpoints for each action (unlike [2]) and do not require any strong thresholds in the framework.

3. RELIABLE-INFERENCE

We formulate our reliable-inference (RI) framework using the "key feature" approach proposed by [7]. The success of inferring world property $\mathcal{P}$ from image feature f in context C can be formulated as the *a posteriori* probability $p(\mathcal{P}|f, C)$. The context C refers to a particular closed-world domain of properties that can occur in some situation. A reliable inference of $\mathcal{P}$ from f makes $p(\mathcal{P}|f, C) \approx 1$ and the probability of an error $p(\neg\mathcal{P}|f, C) \approx 0$. To determine the reliability of f for inferring property $\mathcal{P}$, we form a ratio of these two probabilities

$$R_{post} = \frac{p(\mathcal{P}|f, C)}{p(\neg\mathcal{P}|f, C)} \tag{1}$$

When $R_{post} \gg 1$, the feature f is said to be a highly reliable indicator of property $\mathcal{P}$.

Using Bayes' rule, R_{post} can be separated into the likelihood ratio and the ratio of the priors

$$R_{post} = \frac{p(\mathcal{P}|f, C)}{p(\neg\mathcal{P}|f, C)} = \frac{p(f|\mathcal{P}, C)}{p(f|\neg\mathcal{P}, C)} \cdot \frac{p(\mathcal{P}|C)}{p(\neg\mathcal{P}|C)} \tag{2}$$

A large likelihood ratio indicates that the feature arises consistently with the existence of the world property, but not in its absence. This requirement alone however does not ensure a reliable inference. For if the ratio of priors becomes too small, then R_{post} can become small even in the presence of a large likelihood ratio. Hence a significant context-dependant prior ratio is also required.

3.1. Reliable Action Inference

We are interested in reliable-inference of the action class (world property) given an image pose (feature) of the person. A "key pose" therefore has a feature representation $\mathbf{f}$ (multi-dimensional vector) that can be used to reliably infer a particular action $\mathcal{A}_i$ occurring in context C. We can rewrite Eqn. 1 for the target action $\mathcal{A}_i$ as

$$R_{post} = \frac{p(\mathcal{A}_i|\mathbf{f}, C)}{p(\neg\mathcal{A}_i|\mathbf{f}, C)} = \frac{p(\mathbf{f}|\mathcal{A}_i, C)p(\mathcal{A}_i|C)}{\sum_{j\neq i} p(\mathbf{f}|\mathcal{A}_j, C)p(\mathcal{A}_j|C)} \tag{3}$$

The term $p(\mathbf{f}|\mathcal{A}, C)$ is referred to as the image model, and $p(\mathcal{A}|C)$ is referred to as the action model. The context-dependent reasoning provides a limited domain C of actions for consideration during recognition. For example, if we know the person is traversing the scene, we could possibly limit the context to only locomotory behaviors such as walking and running and greatly reduce the search space of solutions.

To evaluate the R_{post} for $\mathbf{f}$, we model the class likelihoods from training data and select appropriate context-dependent priors.

3.2. Likelihood Modeling

We model the likelihood of feature vector $\mathbf{f}$ appearing from a particular action class $\mathcal{A}_i$ (in a given context) as a Gaus-

sian mixture model

$$p(\mathbf{f}|\mathcal{A}_i) = p(\mathbf{f}|\theta_{\mathcal{A}_i}) = \sum_{k=1}^{K} w_k \cdot g_k(\mathbf{f}|\mu_k, \Sigma_k) \quad (4)$$

where $g_k(\mathbf{f}|\mu_k, \Sigma_k)$ is the likelihood of $\mathbf{f}$ appearing from the k-th Gaussian distribution parameterized by the mean μ_k and covariance Σ_k, with mixture weight w_k. For estimating the parameters $\theta_{\mathcal{A}_i}$, we employ the Expectation Maximization (EM) algorithm [8] that maximizes the class log-likelihood

$$\mathcal{L}(\theta_{\mathcal{A}_i}|\mathbf{f}_1, \cdots, \mathbf{f}_N) = \sum_{n=1}^{N} \log(p(\mathbf{f}_n|\theta_{\mathcal{A}_i})) \quad (5)$$

for all N training examples in class $\mathcal{A}_i$.

Initial values for the means, covariances, and mixture weights in Eqn. 4 can be estimated using K-means clustering of the training samples. To give equal emphasis to each dimension of $\mathbf{f}$, we first whiten [9] the class training data. As the clustering result can vary depending on the seed values (initial means), we repeat the entire EM algorithm multiple times, each time using a K-means clustering result from a different random seed initialization. Finally, we choose the EM mixture model that produces the maximum class log-likelihood (Eqn. 5).

3.2.1. Number of Components

One issue regarding mixture models is the number of clusters/distributions K needed to model the data. Rather than manually selecting an arbitrary K, we automatically select from models of different K, the model that maximizes the Bayesian Information Criterion (BIC) [10].

The BIC for a given model parameterization $\theta_{\mathcal{A}_i}$ is computed as

$$\text{BIC}(\theta_{\mathcal{A}_i}) = 2\mathcal{L}(\theta_{\mathcal{A}_i}|\mathbf{f}_1, \cdots, \mathbf{f}_N) - M\log(N) \quad (6)$$

where M is the number of independent model parameters to be estimated. In our formulation, we have

$$M = K \times (m + \frac{m^2 + m}{2}) + (K - 1) \quad (7)$$

with K distributions, $(m + \frac{m^2+m}{2})$ independent parameters for each mean and covariance ($m = \dim(\mathbf{f})$), and $(K-1)$ independent mixture weights ($\sum w_k = 1$).

Since the class log-likelihood of the mixture model (Eqn. 5) improves when more parameters are added to the model (i.e., larger K), the term $M\log(N)$ is subtracted from (twice) the class log-likelihood in Eqn. 6 to penalize models of increasing complexity. The BIC is maximized in an information theoretic manner for more parsimonious parameterizations.

An iterative split-sample training and validation method is also employed where 50% of the training examples are randomly selected and used by K-means/EM to estimate the model parameters, and the remaining 50% of the samples are used to compute the BIC for that model.

3.3. Reliability Decision

As previously stated, when $R_{post} \gg 1$, $\mathbf{f}$ is a reliable indicator of $\mathcal{A}_i$. But how large does R_{post} need to be for this to happen? In other words, what is the value of the decision threshold $\lambda_{\mathcal{A}_i}$ such that we reliably classify $\mathbf{f}$ as indicating the presence of action $\mathcal{A}_i$? We classify $\mathbf{f}$ as an instance of $\mathcal{A}_i$ when

$$\frac{p(\mathbf{f}|\mathcal{A}_i, C)p(\mathcal{A}_i|C)}{\sum_{j\neq i} p(\mathbf{f}|\mathcal{A}_j, C)p(\mathcal{A}_j|C)} > \lambda_{\mathcal{A}_i} \quad (8)$$

otherwise we make no strong commitment (i.e., choose $\neg\mathcal{A}_i$).

To determine the value of the decision threshold for class $\mathcal{A}_i$, we compute the classification errors for all of the training examples in C using multiple decision thresholds (similar to constructing an ROC curve) and select the threshold $\lambda_{\mathcal{A}_i}$ that produces the lowest two-class R_{post} Bayesian error

$$\begin{aligned} p_\lambda(\text{Error}|C) &= p(\text{class}(\mathbf{f}) = \neg\mathcal{A}_i|\mathcal{A}_i)p(\mathcal{A}_i|C) \quad (9) \\ &\quad + p(\text{class}(\mathbf{f}) = \mathcal{A}_i|\neg\mathcal{A}_i)p(\neg\mathcal{A}_i|C) \\ &= p(\text{class}(\mathbf{f}) = \neg\mathcal{A}_i|\mathcal{A}_i)p(\mathcal{A}_i|C) \quad (10) \\ &\quad + \sum_{j\neq i} p(\text{class}(\mathbf{f}) = \mathcal{A}_i|\mathcal{A}_j)p(\mathcal{A}_j|C) \end{aligned}$$

Alternatively, the error for $\mathcal{A}_i$ could be manually bound and the decision threshold automatically determined to give the lowest error rate possible for the remaining classes $\mathcal{A}_{j\neq i}$.

3.4. Recognition

To perform recognition and determine the action label (if any) for $\mathbf{f}$, we compute the R_{post} of $\mathbf{f}$ for all $\mathcal{A}_i \in C$ and compare each ratio with its decision threshold $\lambda_{\mathcal{A}_i}$. Any class meeting its decision threshold for $\mathbf{f}$ is placed into a clique of potential classifications.

If the clique is empty after examining all classes, then we make no commitment to an action classification (i.e., $\text{class}(\mathbf{f}) = \emptyset$). If the resulting clique contains a single class, then we reliably classify $\mathbf{f}$ to that action. In the event that the clique contains more than one action class (due to independent λ thresholds for each class), we choose the class within the clique having the highest R_{post} (the most reliable inference).

As opposed to ML or MAP approaches that always make a forced-choice classification, RI only makes a class commitment when it is confident enough that the feature vector $\mathbf{f}$ can be reliably used to discriminate the actions.

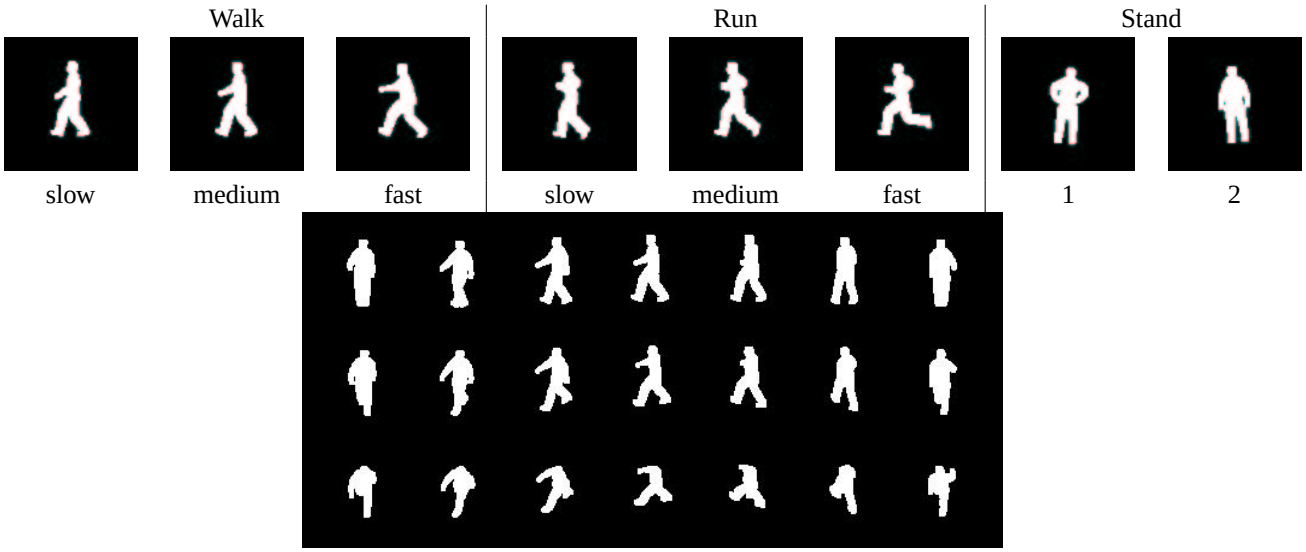


Fig. 1. Example silhouettes for action classes walk, run, and stand. Each class has multiple efforts/styles (top row), and each pose is rendered at 21 different views (bottom image).

4. WALKING, RUNNING, AND STANDING

We selected a context of walking ($\mathcal{W}$), running ($\mathcal{R}$), and standing ($\mathcal{S}$) to evaluate the RI framework for pose-based action recognition. Each action class contains silhouette images of poses at various times, efforts/styles, and views. The walking and running actions were performed at slow, medium, and fast paces to include the natural variations produced at different locomotion speeds [11]. Two common standing poses of hands-on-hips and hands-at-side were also performed, with small movement variations within each style. Example silhouette images are shown in Fig. 1.

4.1. Silhouette Generation

Unless a large number of synchronized cameras at different locations are employed to collect the images, each pose cannot be simultaneously imaged at each viewpoint to conduct consistent view-based evaluations. To address this problem, we used a Vicon-8 motion-capture system and Maya animation software to create a 3-D person model that can be consistently rendered at any desired viewpoint.

We first motion-captured a person performing walking, running, and standing at different efforts/styles. For walking and running, one cycle at each pace was extracted. The motion-capture data was then mapped to a 3-D body model (see Fig. 2), and rendered (orthographic) as a silhouette from multiple viewpoints using OpenGL. Each pose was rendered at 21 different viewpoints separated by 30° horizontal and vertical intervals (see bottom image of Fig. 1).

The silhouettes were rendered small in proportion to the image size ($<$10% of a 100×100 image) and dilated to produce lower-resolution silhouettes comparable to the output of traditional image-based segmentation methods with a remote color or thermal camera (see Fig. 1). The total number of images for classes $\mathcal{W}$, $\mathcal{R}$, and $\mathcal{S}$ were 2184, 1512, and 1974, respectively.

Fig. 2. 3-D body model used to render silhouettes.

4.2. Silhouette Features

We represent each silhouette image with a feature vector of 7 similitude moments [12]. These moments produce excellent global shape descriptors for binary (and grayscale) images in a translation- and scale-invariant manner.

For silhouette image I, its first 7 similitude moments are

given by

$$\eta_{ij} = \frac{\nu_{ij}}{(\nu_{00})^{\frac{i+j}{2}+1}} \tag{11}$$

for orders $2 \leq (i+j) \leq 3$, with the central moments ν_{ij} computed as

$$\nu_{ij} = \sum_x \sum_y (x-\bar{x})^i (y-\bar{y})^j I(x,y) \tag{12}$$

The resulting 7×1 feature vector $\mathbf{f}$ compactly represents the shape of the silhouette image as

$$\mathbf{f} = [\eta_{02}, \eta_{03}, \eta_{11}, \eta_{12}, \eta_{20}, \eta_{21}, \eta_{30}]^T \tag{13}$$

If rotation invariance is also desired, absolute moment invariants [12] could be employed.

5. EXPERIMENTAL EVALUATIONS

We evaluated the RI framework with the action classes $\mathcal{W}$, $\mathcal{R}$, and $\mathcal{S}$ to determine the feasibility of using individual poses for recognition. First we examined the individual R_{post} discrimination results of the actions. Next we compared the RI recognition results to ML and MAP, and also examined the recognition as a function of view-angle. We further analyzed the walking motions using the RI framework to classify the walking pace.

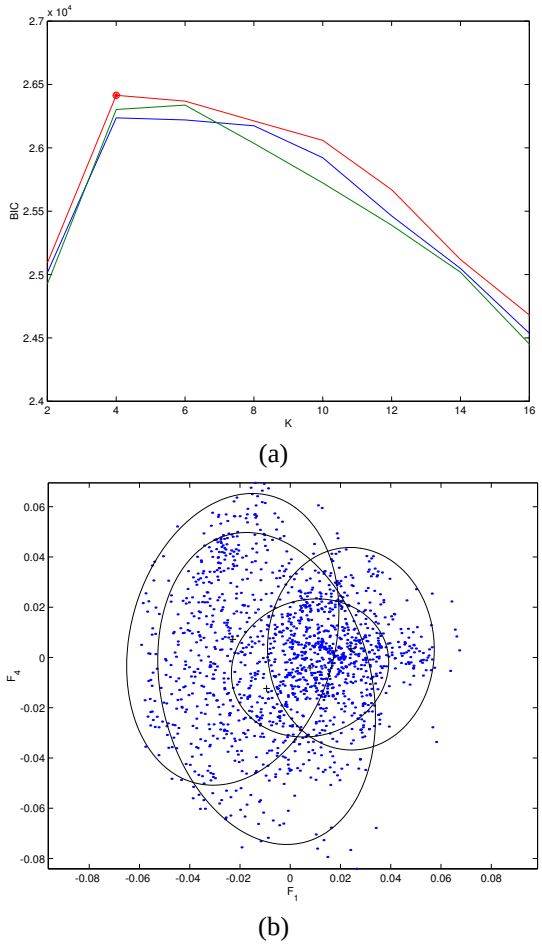
We initially constructed the likelihood mixture model for each class using the approach outlined in Sect. 3.2. For each K under consideration (2–24, in steps of 2), the Kmeans/EM algorithm was repeated 15 times (EM itself was limited to 30 iterations) and the model producing the maximum class log-likelihood was selected as the best model for that K. The best models (one for each K) were then compared using the BIC, and the one having the largest BIC was selected as the optimal model. This entire process was repeated for 3 different split-sample partitions of the class data and the model having the overall largest BIC was selected as the final likelihood model.

In Fig. 3.a, we show the BIC values as a function of K for the running data using three different split-sample iterations. The resulting mixture model corresponding to the maximum BIC (at K=4) is shown in Fig. 3.b.

5.1. Decision Errors in R_{post}

Once the likelihood models were created for each class, the R_{post} decision thresholds were calculated using the method outlined in Sect. 3.3.

We initially employed equal priors: $p(\mathcal{W}|C)$ = $p(\mathcal{R}|C)$ = $p(\mathcal{S}|C) = 1/3$. The R_{post} Bayesian error (Eqn. 10) as a function of λ for running is shown in Fig. 4. The R_{post} errors produced using the optimal decision threshold λ for each class are presented in Table 1.a. We also calculated



(a)

(b)

Fig. 3. Likelihood model for running. (a) BIC values for different K using three split-sample iterations. (b) Mixture model (contour plot at 4σ for whitened moments η_{02}, η_{12}) corresponding to the maximum BIC (at $K = 4$).

the decision thresholds using a different choice of priors: $p(\mathcal{W}|C) = .5$, $p(\mathcal{R}|C) = .2$, and $p(\mathcal{S}|C) = .3$. The resulting R_{post} errors for these priors are presented in Table 1.b for comparison.

The R_{post} Bayesian errors for both sets of priors yield approximately 5% error for walking and running, and only 1% error for standing. This result is encouraging, given only a limited mixture model is used to generalize the features in each class. Therefore the error statistics demonstrate the potential for each class to be reliably distinguished from the remaining classes.

To illustrate the non-uniformity of R_{post} for different poses, we plot in Fig. 5 the ($\mathcal{W}$, $\neg\mathcal{W}$) R_{post} values for a new horizontal side-view ($R_x = 0°$, $R_y = -90°$) three-cycle walking sequence. This plot clearly shows that certain frames are more reliable (having a higher R_{post}) than oth-

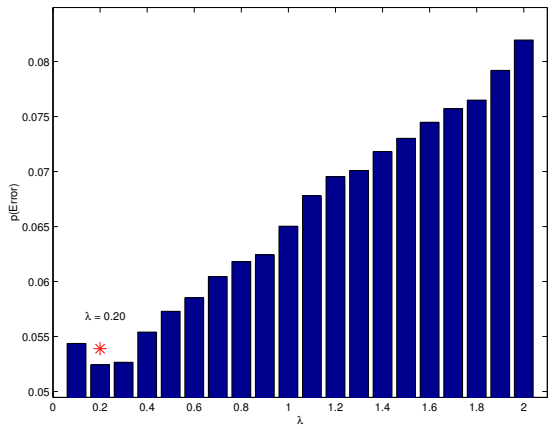


Fig. 4. R_{post} error as a function of λ for $\mathcal{R}$ vs. $\neg\mathcal{R}$ using equal priors. The decision threshold $\lambda = .2$ is marked.

R_{post}	λ	Err$\|\mathcal{W}$	Err$\|\mathcal{R}$	Err$\|\mathcal{S}$	R_{post} error
$\mathcal{W}$, $\neg\mathcal{W}$	8.9	.0847	.0443	.0122	.0471
$\mathcal{R}$, $\neg\mathcal{R}$	0.2	.0600	.0714	.0258	.0524
$\mathcal{S}$, $\neg\mathcal{S}$	0.1	.0069	.0159	.0193	.0140

(a) Equal Priors

R_{post}	λ	Err$\|\mathcal{W}$	Err$\|\mathcal{R}$	Err$\|\mathcal{S}$	R_{post} error
$\mathcal{W}$, $\neg\mathcal{W}$	7.8	.0504	.0847	.0228	.0490
$\mathcal{R}$, $\neg\mathcal{R}$	0.3	.0275	.1336	.0218	.0470
$\mathcal{S}$, $\neg\mathcal{S}$	0.1	.0055	.0159	.0263	.0138

(b) Unequal Priors

Table 1. R_{post} errors corresponding to decision thresholds λ for walking ($\mathcal{W}$), running ($\mathcal{R}$), and standing ($\mathcal{S}$) using (a) equal priors and (b) unequal priors (see text).

ers during the action. We also computed for each class the maximum and minimum R_{post} values for examples at each view. The 5 most reliable and the 5 least reliable poses at different views for each class are shown in Fig. 6.

5.2. Action Recognition

To evaluate the proposed RI recognition method (Sect. 3.4), we compared the RI results to ML and MAP classifications. In Table 2, we present the classification results of RI and ML using equal priors. The overall Bayes error for each method was calculated as

$$\begin{aligned} p(\text{Error}|C) &= p(\text{Error}|\mathcal{W})p(\mathcal{W}|C) \\ &\quad + p(\text{Error}|\mathcal{R})p(\mathcal{R}|C) \\ &\quad + p(\text{Error}|\mathcal{S})p(\mathcal{S}|C) \end{aligned} \tag{14}$$

and yielded 6.31% error for RI and 7.89% error for ML. If we do not consider assignment to $\emptyset$ as an error for RI and

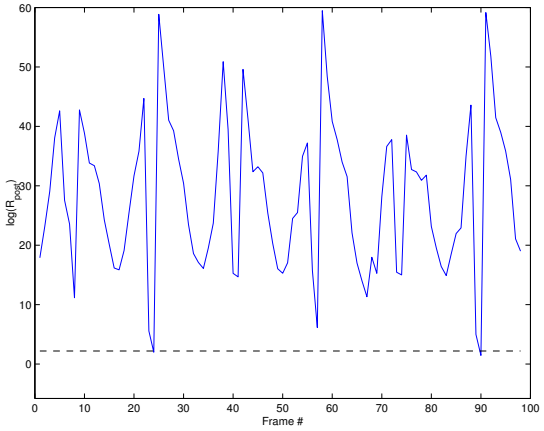


Fig. 5. R_{post} values (log) for a new three-cycle walking sequence.

		Classification				
Input	Method	$\mathcal{W}$	$\mathcal{R}$	$\mathcal{S}$	$\emptyset$	% Error
$\mathcal{W}$	RI	1999	131	14	40	8.47/**6.76**
	ML	2126	53	5	–	2.66
$\mathcal{R}$	RI	67	1399	10	36	7.47/**5.22**
	ML	222	1281	9	–	15.28
$\mathcal{S}$	RI	24	35	1915	0	2.99/**2.99**
	ML	78	35	1861	–	5.72

Table 2. Recognition rates comparing RI and ML classification using equal priors. Errors in bold correspond to using only class-committed examples.

normalize the remaining RI errors by the number of poses actually committed to an action class, the new RI error rate is significantly lowered to 4.99%. In this case, only 1.34% of the frames were unclassified.

The classification results for RI and MAP using the alternate (unequal) priors are presented in Table 3. The Bayes errors were 6.44% for RI and 7.22% for MAP. If the error for RI does not consider the unclassified poses (1.87% unclassified), the Bayes error for RI is reduced to 4.76%.

For both sets of priors, the RI framework produced lower Bayes errors than ML and MAP. With the high FPS available from real-time video, the percentages of unclassified (skipped) poses in each case is insignificant.

5.3. View-Based Discrimination

The previous evaluation computed the classification and error rates using poses at all 21 viewpoints. We next evaluated the recognition capability of RI as a function of the viewpoint to determine which views are most informative toward discrimination of the actions. The Bayes error for

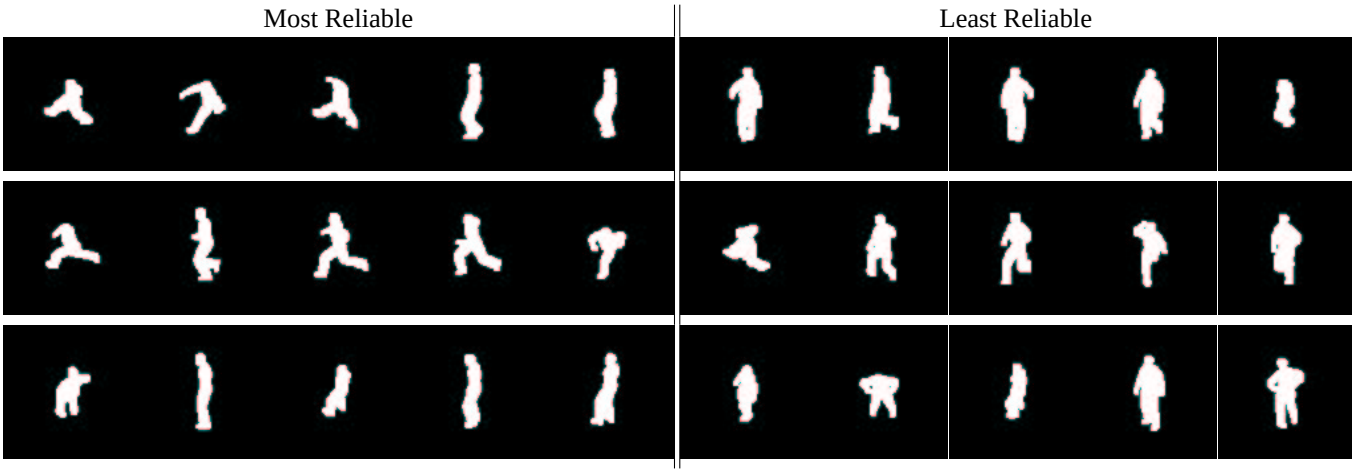


Fig. 6. The five most reliable poses and five least reliable poses in terms of R_{post} at different views. Decreasing pose reliability is ordered from left-to-right in each row. Top row: Walking, Middle row: Running, Bottom row: Standing.

		Classification				
Input	Method	$\mathcal{W}$	$\mathcal{R}$	$\mathcal{S}$	$\emptyset$	% Error
$\mathcal{W}$	RI	2074	60	10	40	5.04/**3.26**
	MAP	2148	32	4	–	1.65
$\mathcal{R}$	RI	128	1305	15	64	13.69/**9.88**
	MAP	337	1164	11	–	23.02
$\mathcal{S}$	RI	45	31	1896	2	3.95/**3.85**
	MAP	87	31	1856	–	5.98

Table 3. Recognition rates comparing RI and MAP classification using unequal priors.

		R_y						
		$0°$	$-30°$	$-60°$	$-90°$	$-120°$	$-150°$	$-180°$
R_x	$0°$	.01	.10	.01	.01	.02	.12	.04
	$30°$	.06	.10	.04	.01	.00	.15	.05
	$60°$	.02	.02	.07	.12	.04	.12	.22

Table 4. Bayes error for walking, running, and standing at each view.

the poses at each of the 21 views is presented in Table 4. As expected, the best views for recognition were located near the side ($R_y = -90°$) at nearly horizontal views. Interestingly, a downward looking view from behind the person produced the largest error (22%).

5.4. Identifying Walking Pace

To further demonstrate the RI method in terms of identifying confusing poses, we examined the pose differences in the slow, medium, and fast walking paces (see Fig. 1) at multiple views. As these walking efforts are very similar in appearance, we expect the RI method to identify several poses that are too confusing to classify.

R_{post}	λ	R_{post} error
$\mathcal{W}_{\text{slow}}$, $\neg\mathcal{W}_{\text{slow}}$	5.3	.2589
$\mathcal{W}_{\text{med}}$, $\neg\mathcal{W}_{\text{med}}$	1.6	.3173
$\mathcal{W}_{\text{fast}}$, $\neg\mathcal{W}_{\text{fast}}$	4.8	.2003

Table 5. R_{post} errors corresponding to decision thresholds λ for slow, medium, and fast walking paces using equal priors.

The likelihood mixture model for each walking pace was estimated using the approach in Sect. 3.2. The R_{post} errors for the walking paces using equal priors are reported in Table 5. As expected the R_{post} discrimination errors are quite large (20–32%). The most reliable and least reliable fronto-parallel pose for each pace are shown in Fig. 7. The most reliable poses at this view appear to capture different stride extensions.

In Table 6, we present a comparison of the RI and ML classification results. For each walking pace, several poses were deemed unreliable by RI and were therefore placed in the $\emptyset$ category. The RI Bayes error was 59.54% and the ML Bayes error was 41.71%. Without consideration of the unclassified poses (42% unclassified), the error for RI was 32.00%.

Though the RI approach did not classify 42% of the poses, the method is still applicable given that there are typically 30–40 frames during a single walk cycle with 30 FPS video. Therefore, to reduce the classification error rate, the RI approach is still desired over ML.

		Classification				
Input	Method	Slow	Med	Fast	$\emptyset$	% Error
Slow	RI	370	77	13	380	55.95/**19.57**
	ML	620	129	91	–	26.19
Med	RI	120	169	87	338	76.33/**55.05**
	ML	273	232	209	–	67.51
Fast	RI	31	61	338	200	46.35/**21.40**
	ML	92	106	432	–	31.43

Table 6. Recognition rates comparing RI and ML classification of slow, medium, and fast walking using equal priors.

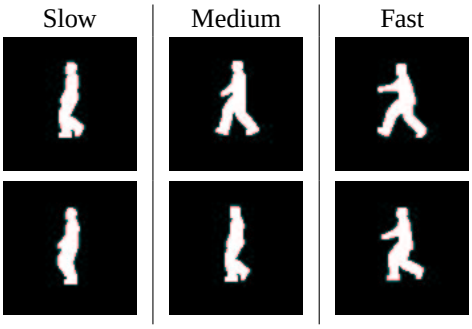


Fig. 7. The most reliable (top row) and least reliable (bottom row) fronto-parallel poses for slow, medium, and fast walking.

6. SUMMARY AND CONCLUSIONS

We presented a reliable-inference approach to evaluate the feasibility of action recognition from single poses. The approach is formulated in a probabilistic framework that verifies the reliability of inference from a pose before committing to any action classification. To determine that a pose is a reliable indicator of action $\mathcal{A}_i$, we form the *a posteriori* probability ratio R_{post} for classes $\mathcal{A}_i$ and $\neg\mathcal{A}_i$, and check that it is above a minimum Bayesian error threshold derived from the training data. To model the class likelihoods, we outlined an EM-based Gaussian mixture-model technique using the Bayesian Information Criterion to automatically determine the optimal number of mixture components.

For recognition of a given pose, we select the class having the largest valid R_{post}. If no class has a valid R_{post} for the pose, then the system does not commit to any action classification. The recognition results for walking, running, and standing at multiple views from individual poses showed encouraging results with approximately 5% Bayes error for class-committed poses (ML=8%, MAP=7%).

In future work, we plan to train and evaluate the system with multiple actions of several people in outdoor scenes. We are also considering the use of MHIs [13] for short-duration action modeling. As the moment features are global descriptors and sensitive to major occlusions, we plan to investigate local part-based feature representations. We additionally are interested in evaluating the RI results in a man-machine comparison.

7. REFERENCES

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

[2] M. Oren, C. Papageorgiour, P. Sinha, E. Osuma, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. Comp. Vis. and Pattern Rec.* IEEE, 1997, pp. 193–199.

[3] D. Gavrila, "Pedestrian detection from a moving vehicle," in *Proc. Euro. Conf. Comp. Vis.*, 2000, pp. 37–49.

[4] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," in *Proc. Wkshp. Applications of Comp. Vis.*, 1998.

[5] A. Baumberg and D. Hogg, "Learning flexible models from image sequences," in *Proc. Euro. Conf. Comp. Vis.*, 1994, pp. 299–308.

[6] R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts," in *Proc. Comp. Vis. and Pattern Rec.* IEEE, 2000, pp. 721–727.

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

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[9] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, New York, 2001.

[10] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] J. Davis and S. Taylor, "Analysis and recognition of walking movements," in *Proc. Int. Conf. Pat. Rec.*, 2002, pp. 315–318.

[12] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Information Theory*, vol. IT-8, no. 2, pp. 179–187, 1962.

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