A Reliable-Inference Framework for Recognition of Human Actions*

James W. Davis Ambrish Tyagi
Dept. of Computer and Information Science
Ohio State University
Columbus, OH 43210 USA
{jwdavis,tyagia}@cis.ohio-state.edu

Abstract

We present an action recognition method based on the concept of reliable inference. Our approach is formulated in a probabilistic framework using posterior class ratios to verify the saliency of an input before committing to any action classification. The framework is evaluated in the context of walking, running, and standing at multiple views and compared to ML and MAP approaches. Results examining individual silhouette images with the framework demonstrate that these actions can be reliably discriminated while discounting confusing images.

1. Introduction

Advanced video surveillance systems will require the capability to detect the presence of people, track their movements, and recognize their behaviors and actions. Typically, analysis over *several* frames is employed to construct representations for recognition (e.g., matching trajectories or detecting characteristic motion patterns [1]). But how reliably could a system perform when limited to analysis of only one frame, or two frames, or three frames, etc?

Clearly, reliable recognition of basic activities from the smallest number of video frames would be advantageous to automatic video-based surveillance, especially for systems having limited computational processing time scheduled per camera or for systems employing time-lapse recording/processing. Also consider small unmanned aerial vehicles (UAVs) or mobile robotic platforms equipped with video cameras. These systems have a constantly changing view area, and therefore immediate decisions about the activity in the scene are desirable. Even if longer duration video is available in some systems, rapid action detection may be particularly helpful in bootstrapping more sophisticated action-specific tracking or recognition approaches.

However, as the number of frames is reduced, there will be more actions confused during recognition. To evaluate different short-term durations, an appropriate framework capable of properly handling inconclusive information is therefore desired over a forced-choice classification method which can produce noisy results. We present a method that automatically identifies the confusing information and removes it from consideration during classification.

Our approach first examines an input with a series of *a posteriori* class comparisons to evaluate its discrimination reliability. Only when the input is deemed "good enough" for discrimination between the possible actions does a classification take place. This approach is particularly favorable when there is a high cost for making errors and low (or no) cost for passively waiting for more information to arrive (advantageous with real-time video). 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.

We evaluate the proposed reliable-inference framework using the task of discriminating walking, running, and standing at multiple viewpoints. In this paper, we push the classification task to the extreme and only consider input of a single image (though the approach is clearly applicable to other sequence-based analysis methods [3]). As single frames from different views have more classification ambiguity, rather than if multiple frames are considered, this domain is a particularly good experimental testbed for the reliable-inference approach.

We present results examining the framework with the walking, running, and standing actions, and show that low Bayes error rates can be achieved. We also make comparisons to alternative ML and MAP approaches, and examine the discrimination ability as a function of viewpoint to determine the best camera locations to recognize the actions. To further illustrate the detection and elimination of confusing poses, we present results discriminating subtle changes

*Appears in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Miami, Florida, July 21-22, 2003, pp. 169-176.

in walking pace (slow, medium, fast).

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

2. Related Work

In [11], 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 was used in [7] to also detect pedestrians. The template hierarchy was constructed automatically from examples using refinement clustering of images into prototypes. 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 an RBF classifier.

For discriminating humans and vehicles, two simple properties (dispersedness, area) were used by [10] 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 [2] 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 [12], 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 application of the reliable-inference framework similarly employs an EM-based clustering of silhouette poses using moment features as in [12], 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 information during immediate decision-making, rather than necessarily requiring temporal consistency before classification. We also examine multiple viewpoints for each action and do not require any strong manual thresholds in the framework.

3. Reliable-Inference Framework

We formulate our reliable-inference (RI) framework using the "key feature" approach proposed by [9]. 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

In this paper, we are interested in reliable-inference of the action class (world property) given an image (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 to reduce the search space of solutions.

To evaluate the R_{post} for $\mathbf{f}$, we first 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 Gaussian mixture model

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

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 [5] that maximizes the class log-likelihood

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

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 (after whitening [6] to give equal emphasis to each dimension of $\mathbf{f}$). 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) [13].

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

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

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

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

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 used 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 evaluation of 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 $\mathcal{A}_i$? Formally, we classify $\mathbf{f}$ as an instance of $\mathcal{A}_i$ when

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

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

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

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

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 own 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., 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.

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. Each action class contains silhouette images of poses at various times, efforts/styles, and views. 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 for our experiments, we used a Vicon-8 motion-capture system and Maya animation software to create a 3-D person model to render each action (1 cycle) at multiple viewpoints.

The walking and running actions were performed at slow, medium, and fast paces to include the natural variations produced at different locomotion speeds [4]. 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. Each pose was rendered at 21 different viewpoints separated by 30° horizontal and vertical intervals (see bottom image of Fig. 1). The total number of images for classes $\mathcal{W}$, $\mathcal{R}$, and $\mathcal{S}$ were 2184, 1512, and 1974, respectively.

4.1. Silhouette Features

We represented each silhouette image with a feature vector of 7 similitude moments [8]. These moments produce excellent global shape descriptors for binary (and grayscale) images in a translation- and scale-invariant manner. If rotation invariance is also desired, absolute moment invariants [8] could be employed.

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

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

We make no particular claim that these are the optimal features, and many other types of feature descriptors could have been used to represent the silhouette shapes.

5. Experimental Evaluations

First we examined the individual R_{post} discrimination results of the actions using all of the silhouette images. 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 K-means/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. 2.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. 2.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. 3. The R_{post} errors produced using the optimal decision threshold λ for

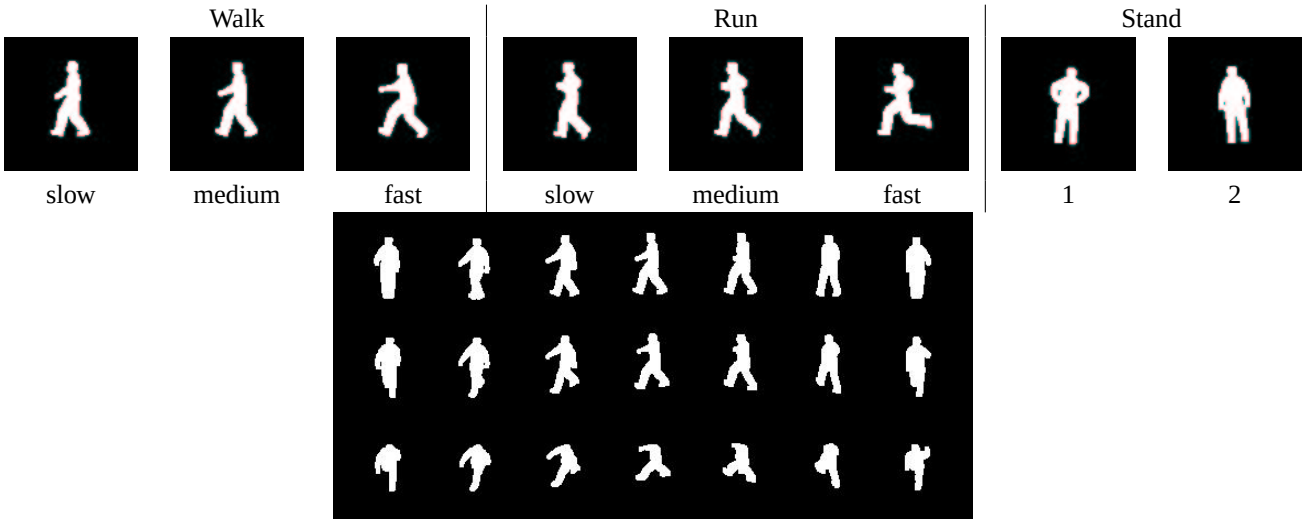


Figure 1. Example silhouettes for action classes walk, run, and stand. Each class has multiple efforts/styles (top row), and each pose is rendered at 21 different views (bottom image).

each class are presented in Table 1.a. We also calculated 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 images over time, we plot in Fig. 4 the $(\mathcal{W}, \neg\mathcal{W})$ R_{post} values for a non-training 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 others during the action. We also computed for each class the maximum and minimum R_{post} values for examples across all views. The most reliable and least reliable pose for each class are shown in Fig. 5.

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

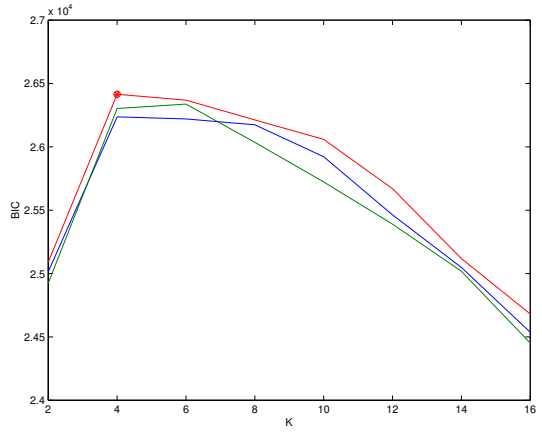
Table 1. R_{post} errors corresponding to decision thresholds λ for walking ($\mathcal{W}$), running ($\mathcal{R}$), and standing ($\mathcal{S}$) using (a) equal priors and (b) unequal priors (see text).

5.2. Recognition

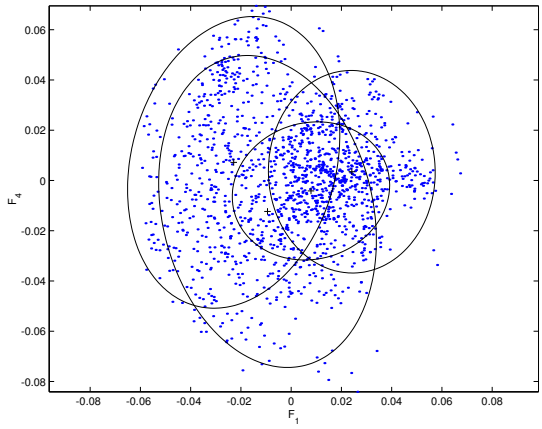
To evaluate the proposed RI recognition method (Sect. 3.4), we compared the RI results to ML and MAP classifications. In Table 2, we present the classification results of RI and ML using equal priors (MAP is equivalent to ML when using equal priors). The overall Bayes error for each method was calculated 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 normalize the remaining RI errors by the number of images actually committed to an action class, the new RI error rate is lowered to 4.99%. In this case, 76 frames (1.34% of the total) were unclassified.



(a)



(b)

Figure 2. Likelihood model for running. (a) BIC values for different K using three split-sample iterations. (b) Mixture model (contour plot at 4σ) corresponding to the maximum BIC (at $K = 4$).

The classification results for RI vs. MAP using the alternate (unequal) priors are presented in Table 3. The Bayes errors were 6.44% for RI and 7.22% for MAP. Again, if we do not consider the unclassified images (106 frames,1.87%), 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) frames in each case is insignificant. Though the improvements in error rates were fairly small in these examples, they nonetheless demonstrate that the method is capable of achieving a better performance and identifying confusing information (which may produce more significant improvements in other cases).

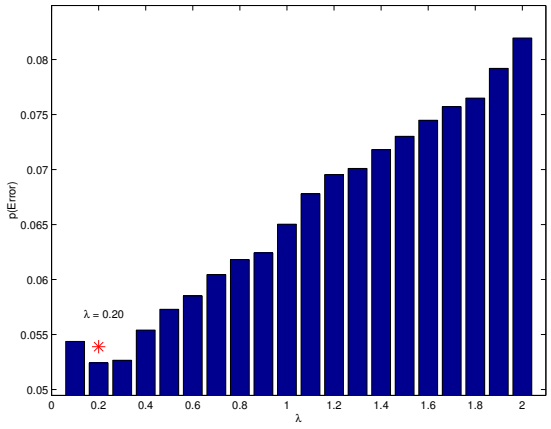


Figure 3. R_{post} error as a function of λ for $\mathcal{R}$ vs. $\neg\mathcal{R}$ using equal priors.

		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.

5.3. View-Based Discrimination

The previous evaluation computed classification and error rates using 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 given actions. The Bayes error for the images 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 mostly 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 images, we examined the differences in the slow, medium, and fast walking paces (see Fig. 1) at

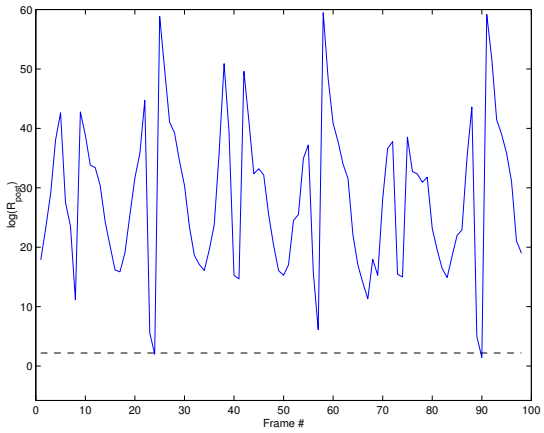


Figure 4. R_{post} values (log) for a new three-cycle walking sequence.

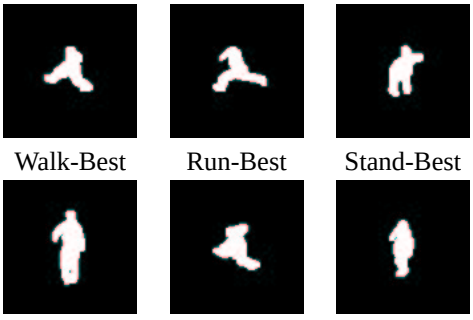


Figure 5. The most reliable and least reliable poses in terms of R_{post}.

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.

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 side-view pose for each pace are shown in Fig. 6, where 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 this data. 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

		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.

error for RI was reduced to 32.00%.

Though the RI approach did not classify 42% of the images, the method is still applicable given that there are typically 30–40 frames during a single walk cycle with 30 FPS video (thus more than half of the frames per walk cycle are reliably classifiable on average).

6. Summary and Conclusions

We presented a method for reliable inference of human actions. The approach is formulated in a probabilistic framework that first verifies the reliability of inference of an input before committing to any action classification. To determine that an input 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, we select the class having the largest valid R_{post}. If no class has a valid R_{post}, then the system does not commit to any action classification. The recognition results examining single frames of walking, running, and standing at multiple views showed encouraging results with approximately 5% Bayes error for class-committed poses (ML=8%, MAP=7%).

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.

		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.

In future work, we plan to train and evaluate the system with multiple actions of several people in outdoor scenes. We are currently developing a night-vision surveillance system using thermal cameras that produce images amenable to our framework (See Fig. 7). We also plan to investigate local part-based feature representations to compare with the global moment descriptors. As the framework is not inherently constrained to use only single images as input, our next step is to evaluate the approach with multiple frames using Motion History Images (MHIs) [3] for short-duration action modeling.

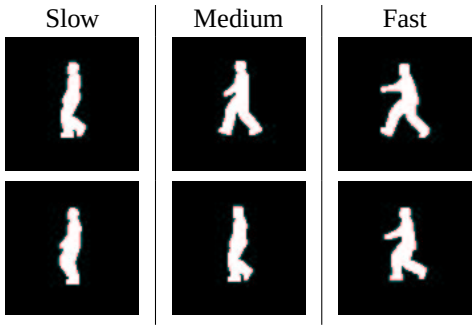


Figure 6. The most reliable (top row) and least reliable (bottom row) side-view poses for slow, medium, and fast walking.

Figure 7. Example thermal image showing people in an outdoor environment.

References

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

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

[3] 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.

[4] J. Davis and S. Taylor. Analysis and recognition of walking movements. In *Proc. Int. Conf. Pat. Rec.*, pages 315–318, 2002.

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

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

[7] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. European Conf. Comp. Vis.*, pages 37–49, 2000.

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

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

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

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

[12] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Proc. Comp. Vis. and Pattern Rec.*, pages 721–727. IEEE, 2000.

[13] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.