Sequential Reliable-Inference for Rapid Detection of Human Actions*

James W. Davis
Dept. of Computer Science and Engineering
Ohio State University
Columbus, OH 43210 USA
jwdavis@cse.ohio-state.edu

Abstract

We present a probabilistic reliable-inference framework to address the issue of rapid-and-reliable detection of human actions. The approach determines the shortest video exposure needed for low-latency recognition by sequentially evaluating a series of posterior class ratios to find the earliest reliable decision point. Results are presented for a set of people walking, running, and standing at different styles and multiple viewpoints, and compared to an alternative ML approach.

1. Introduction

A fundamental question regarding activity recognition is how much time is actually necessary to identify common human actions? In the extreme case, people can easily recognize several different actions from looking at just a single picture – one need only flip through the pages of a newspaper or magazine to make this point. But when we see an action at some random time (offset) during the movement or at a non-canonical view, a longer exposure may be required to reliably distinguish it from other actions. Once the action has been detected, a continual verification process of the selected action likelihood could be used to retain the classification label over the remainder of the sequence (or announce when the action has changed).

From an applied standpoint, rapid activity recognition is advantageous to automatic video-based surveillance for systems having limited processing time scheduled per camera or for systems employing time-lapse recording. UAVs and mobile robotic platforms with quickly changing camera views similarly require immediate decisions about the current scene activity. Rapid detection may be particularly useful as an initialization stage for bootstrapping more sophisticated action-specific tracking or recognition approaches.

In an early perceptual study [12], it was shown that only a brief exposure time (200 milliseconds) of moving point-lights (attached to the joints of a moving person) was required for people to reliably distinguish a set different actions (variations of walking, running, and jumping jacks). This work helps support the notion that long exposures or repetitions (multiple cycles) of an action may not necessarily be needed for reliable recognition.

We present an efficient probabilistic decision framework for rapid action classification based on the concept of sequential reliable-inference. Instead of forcing a non-reliable classification for constant-duration short exposures, we examine the reliability of the shortest time exposure first and extend it (in time) only until a reliable classification of a particular action can be made. The sequential inference framework and the decision thresholds for each exposure are automatically computed from training examples. The approach is designed to handle all possible starting offsets of the activities (not limited to start at the first frame). We report results evaluating the concept of the framework for distinguishing walking, running, and standing actions at different views, and compare the approach with an alternative maximum-likelihood decision method.

We begin with a review of related single- and multi-frame recognition approaches in Sect. 2. Next we give an introduction to reliable-inference in Sect. 3. We then present the proposed sequential reliable-inference approach in Sect. 4, including algorithm details for temporal binding, feature selection, and likelihood modeling. We present experimental results in Sect. 5. Lastly, we conclude with a summary in Sect. 6.

2. Related Work

Many different approaches related to action recognition have been proposed and are concisely reviewed in [1, 8, 22]. We briefly mention only a few approaches within the single- and multiple-frame analysis domains.

Robust methods have been proposed for identifying walking humans (pedestrians) in single images, including

*Appears in *IEEE Workshop on Detection and Recognition of Events in Video*, Washington DC, July 2, 2004.

the use of wavelets [18], coarse-to-fine matching [9], and point-distribution models [2]. Using two frames, the AdaBoosting technique was employed in [21]. Two simple region properties (dispersedness, area) were used in [13] to classify regions (selected from image differencing), followed by a temporal consistency measure (histogram) to choose the most frequent action label assigned over several frames.

Dynamic action recognition from sequences generally include methods using analysis of trajectories or templates. For periodic actions such as walking and running, trajectory-based approaches for single and multiple cycle exposures include frequency-based Fourier methods [14], feature-based properties (e.g., stride) [6], spatio-temporal patterns [17], and HMMs [4]. With no part tracking, a different approach is to use spatio-temporal templates derived from the image sequence directly, including generic layered templates [3] and other periodic template representations [19, 15, 5, 16].

We approach the action recognition problem differently than the above work in that we seek to determine the *shortest-duration* video exposures needed to achieve reliable recognition with minimal latency. We note that we could potentially employ some of the above approaches to represent video exposures within our decision framework (we currently use [3]).

3. Reliable-Inference

We formulate our reliable-inference (RI) framework using the "key feature" proposal of [11]. It states that the success of inferring a world property $\mathcal{P}$ from an image feature f in context C can be formulated as the *a posteriori* probability $p(\mathcal{P}|f, C)$, where the context C refers to a particular closed-world domain. A reliable-inference of $\mathcal{P}$ from f makes $p(\mathcal{P}|f, C)$ large, and the probability of making an error $p(\neg\mathcal{P}|f, C) \approx 0$. A single reliability measurement of f for inferring $\mathcal{P}$ is formed by the ratio of these two posterior 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 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} becomes small even in the presence of a large likelihood ratio. Hence a significant context-dependant prior ratio is also required.

In our domain of action recognition, we consider $\mathbf{f}$ to be a multi-dimensional feature vector for a given video exposure of an action sequence, and define $\{\mathcal{A}\}_C$ to be the set of possible actions (world properties) in the context C. We can rewrite the R_{post} in Eqn. 2 for a specific action $\mathcal{A}_i \in \{\mathcal{A}\}_C$ as

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

For a given reliability threshold $T_{\mathcal{A}_i}$, we will say that $\mathbf{f}$ reliably indicates the presence of action $\mathcal{A}_i$ if its $R_{post} > T_{\mathcal{A}_i}$.

4. Sequential Reliable-Inference

To reliably recognize actions within the smallest video exposure time (from any temporal offset), the preceding reliable-inference formulation must be extended to efficiently accommodate sequential decisions at several different exposure lengths. Additionally, we must be able to select an appropriate R_{post} threshold for each action class at each exposure length.

We present a multi-level sequential RI method that automatically determines an appropriate R_{post} threshold for each action exposure from negative examples and links the training failures (unreliable frames) from one level/exposure to the training input of the next (longer) exposure level. For recognition, the framework continually incorporates a new video frame into a longer video exposure only until a valid R_{post} for an action class is found. The approach is depicted in Fig. 1.

4.1. Multi-Level Analysis

The sequential RI approach begins with determining the appropriate R_{post} thresholds for each action class using only a single video frame. Here all training sequences for an action are grouped as a collection of individual/distinct frames. A feature vector is computed for each frame, and the likelihood distribution is estimated for each class.

For action class $\mathcal{A}_i$, we compute the maximum R_{post} of all the negative ($\neg\mathcal{A}_i$) single-frame examples as

$$T_{neg} = \max \frac{p(\mathbf{f}|\mathcal{A}_i, C)p(\mathcal{A}_i|C)}{\sum_{j\neq i} p(\mathbf{f}|\mathcal{A}_j, C)p(\mathcal{A}_j|C)}, \forall \mathbf{f} \in \neg\mathcal{A}_i \tag{4}$$

Next we find the minimum $R_{post} > T_{neg}$ of the positive class ($\mathcal{A}_i$) examples with

$$T_{pos} = \min \frac{p(\mathbf{f}|\mathcal{A}_i, C)p(\mathcal{A}_i|C)}{\sum_{j\neq i} p(\mathbf{f}|\mathcal{A}_j, C)p(\mathcal{A}_j|C)} > T_{neg}, \forall \mathbf{f} \in \mathcal{A}_i \tag{5}$$

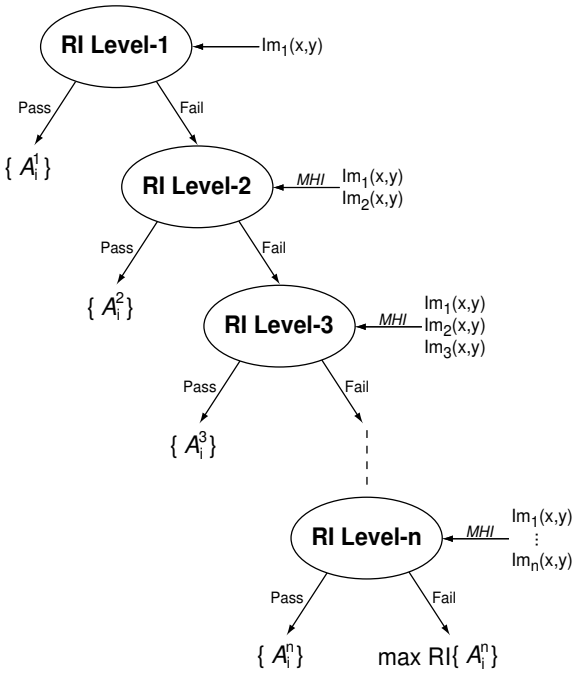


Figure 1: Sequential reliable-inference.

and set the final R_{post} threshold for class $\mathcal{A}_i$ as

$$T_{\mathcal{A}_i} = (T_{neg} + T_{pos})/2 \tag{6}$$

The use of negative examples enforces that no early classification errors occur within the training set, as additional video frames would likely help to alleviate the ambiguities.

The positive class examples passing its R_{post} threshold are deemed as reliable single-frame indicators of that action (i.e., key poses). Those training examples that fail their respective R_{post} threshold are collected and passed on to the next level (two-frame RI). Thus less examples are evaluated in the next exposure length.

At the second level, we form a new training set of all the unreliable (un-classified) frames from the previous single-frame level. Each example is extended in time to include the next video frame within its respective sequence to form a training set of tuple frames. Any example that happens to be the last frame in a training sequence (having no successor frame) is discarded. A feature vector for each tuple is then computed, and the two-frame likelihood distribution is estimated for each class. The R_{post} thresholds for this level/exposure are calculated as in the first level, but using only the tuple examples at this level. The positive example failures for each class are then passed on to the next three-frame exposure level for analysis.

The process of computing the R_{post} thresholds and identifying class failures at each level is continued until a stopping criteria is met. An action class terminates at a particular level when its likelihood distribution cannot be calculated due to an insufficient number of training examples. The hierarchy itself terminates when only one action class (or none) remains (R_{post} requires a minimum of two classes for comparison).

As some action classes may have reliably-inferred proportionately more training examples than other classes at a particular level (thus passing fewer examples into the next level), the action priors should be adjusted to reflect the occurrence of the actions in the next level. Intuitively, if less examples for a class are continued into the next level, there is less of a chance of seeing that class as compared with the previous level. For level $l > 1$, we weight the priors from the previous level ($l - 1$) by the fraction of examples ω in each class that are continued

$$p_l(\mathcal{A}_i|C) = \frac{\omega_i \cdot p_{l-1}(\mathcal{A}_i|C)}{\sum_j \omega_j \cdot p_{l-1}(\mathcal{A}_j|C)} \tag{7}$$

If any class is terminated in level $l - 1$, the remaining class priors are normalized before applying Eqn. 7.

An advantage to this sequential decision framework, in terms of training, is that the data is systematically reduced at each level of the hierarchy (i.e., the deeper in the tree, the lesser amount of training data being used), and therefore less computation and potentially better class separability can be achieved.

4.2. Recognition

For recognition, the first digitized video frame is examined at the initial level of the hierarchy (single-frame exposure). If the frame is found to be a reliable indicator of a particular action (above an action's R_{post} threshold), we assign that class label and exit the search. If instead the frame is deemed unreliable (below every R_{post} threshold at that level), we include the next video frame to make a two-frame exposure and examine the R_{post} values in the second level looking for a reliable match. This process is continued down through the levels (increasing the exposure) until either 1) a reliable-inference is found, 2) no more exposure levels exist, 3) the final level contains one class, or 4) no additional video frames are available to extend the exposure.

In case 2, when no match is found in the last level (containing >1 actions), we select the action at this level having the largest R_{post} (most reliable). If only one class remains at the bottom level (case 3), then we choose this action by default (only class available in the context). Lastly (case 4), if an additional video frame for the next level is not available (e.g., due to occlusion, exit of scene, etc.), we choose the action from the previous level with the largest R_{post}.

4.3. Technical Details

4.3.1. Temporal Binding

We employ a temporal binding mechanism to extract a small feature vector for multi-frame video exposures (rather than concatenating the features for each image in the exposure). We use the Motion History Image (MHI) representation [3] to collapse the video exposure into a single 2-D template that perceptually captures the essence of the movement while retaining much of its temporal structure. MHIs are well-suited to representing short-duration movements (where complex tracking is not required).

An MHI is generated by layering successive images of a person using a replacement-and-decay operator. The MHI at frame t (of a δ-frame exposure) is updated as

$$\text{MHI}_{\delta}^{t}(x,y) = \begin{cases} t/\delta & \text{if } \Psi(I^t(x,y)) \neq 0 \\ \text{MHI}_{\delta}^{t-1}(x,y) & \text{otherwise} \end{cases} \tag{8}$$

where each pixel (x,y) in the new MHI (at time t) is marked with a normalized timestamp t/δ if the function Ψ signals presence of the person (e.g., silhouette, motion, or skin color) in the current video image $I^t(x,y)$. The MHI pixels where $\Psi(I^t(x,y)) = 0$ remain unchanged from the previous update. This function is called for every new video frame analyzed in the δ-frame exposure.

4.3.2. Features

We represent each MHI with a feature vector of 7 similitude moments that exhibit scale and translation invariance [10]. These moments produce excellent global shape descriptors for grayscale and binary images and have previously been demonstrated with MHIs [3]. For a given MHI, its first 7 similitude moments are computed as

$$\eta_{ij} = \frac{\sum_x \sum_y (x-\bar{x})^i (y-\bar{y})^j \text{MHI}(x,y)}{\left[\sum_x \sum_y \text{MHI}(x,y)\right]^{\frac{i+j}{2}+1}} \tag{9}$$

for orders $2 \leq (i+j) \leq 3$, resulting in a compact 7×1 feature vector $\mathbf{f} = [\eta_{02}, \eta_{03}, \eta_{11}, \eta_{12}, \eta_{20}, \eta_{21}, \eta_{30}]^T$. If rotation invariance is also desired, absolute moment invariants [10] could be employed. We make no particular claim that these are the optimal features (many other types of feature descriptors could instead be employed).

4.3.3. Likelihood Modeling

To compute the R_{post} for an action class, a likelihood distribution for each action class is required. We model the likelihood of feature vector $\mathbf{f}$ appearing from a particular action class $\mathcal{A}_i$ as a Gaussian mixture model

$$p(\mathbf{f}|\mathcal{A}_i, C) = p(\mathbf{f}|\theta_{\mathcal{A}_i}) = \sum_{k=1}^{K} w_k \cdot g_k(\mathbf{f}|\mu_k, \Sigma_k) \tag{10}$$

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 [7] that maximizes the class log-likelihood

$$\mathcal{L}(\theta_{\mathcal{A}_i}|\mathbf{f}_1, \cdots, \mathbf{f}_N) = \sum_{n=1}^{N} \log(p(\mathbf{f}_n|\theta_{\mathcal{A}_i})) \tag{11}$$

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

Initial values for the means, covariances, and mixture weights in Eqn. 10 can be estimated from a pre-clustering of the training data using K-means. One issue regarding mixture models is the number of distributions K needed to model the data. Rather than manually choosing an arbitrary K, we employ a Minimum Description Length technique to automatically select, from a set of models (each model constructed using a different K), the model that maximizes the Bayesian Information Criterion (BIC) [20].

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

$$\text{BIC}(\theta_{\mathcal{A}_i}) = 2\mathcal{L}(\theta_{\mathcal{A}_i}|\mathbf{f}_1, \cdots, \mathbf{f}_N) - M\log(N) \tag{12}$$

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

$$M = K \times (m + \frac{m^2+m}{2}) + (K-1) \tag{13}$$

with K distributions, $(m + \frac{m^2+m}{2})$ independent parameters for each mean and covariance ($m = \dim(\mathbf{f}) = 7$), and $(K-1)$ independent mixture weights ($\sum w_k = 1$).

Since the class log-likelihood of the mixture model (Eqn. 11) improves when more parameters are added to the model (i.e., using a larger K), the term $M\log(N)$ is subtracted from (twice) the class log-likelihood in Eqn. 12 to penalize models of increasing complexity. The BIC is maximized in an information theoretic manner for more parsimonious parameterizations (to find the optimal K).

5. Experimental Results

To demonstrate the sequential RI framework, we selected common activities of people walking, running, and standing at different styles and multiple viewpoints. We report results on both the multi-level RI construction and the corresponding recognition approach. We also compare the results to a sequential classification using maximum-likelihood (ML) instead of RI to determine the class thresholds at each level.

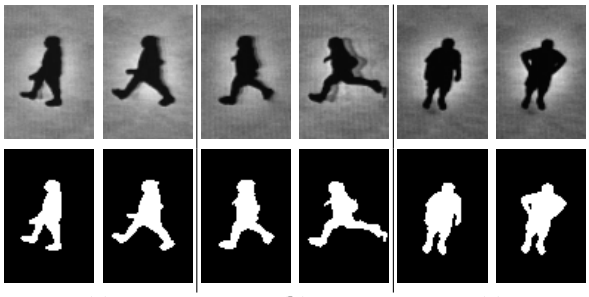


Figure 2: Example thermal images and silhouettes for (a) walk slow/fast, (b) run slow/fast, and (c) stand hands-side/hands-hips.

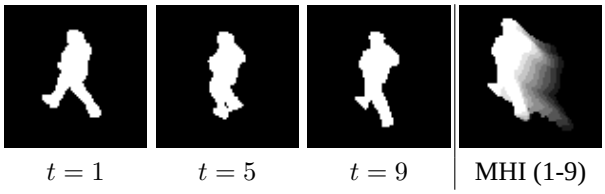


Figure 3: MHI for 9-frame running exposure.

5.1. Action Context

We recorded three people walking, running, and standing from eight different viewpoints using a FLIR (thermal) video surveillance camera.

The walking and running actions were performed at slow and fast paces to include the natural variations produced at different speeds. One cycle at each pace was manually segmented from the video. For standing, two common poses of hands-at-side and hands-on-hips were performed. The total number of images of walk, run, and stand were 1416 (avg. 30 frames/seq), 1022 (avg. 21 frames/seq), and 516 (avg. 11 frames/seq), respectively. As the variation within the static stand styles was minimal, we retained less frames than the other classes (though sufficient to model the likelihood distributions at multiple levels).

A simple background subtraction technique suited to thermal imagery was employed to extract the person silhouettes (see Fig. 2). Half of the silhouette images were randomly selected as starting (initial offset) frames for training and the remaining examples were used as starting frames for testing the recognition approach.

5.2. Training

Using the training method outlined in Sect. 4.1, we constructed the RI hierarchy using the walk, run, and stand training examples. The actions were initially assigned equal priors. The MHI temporal binding and moment features of

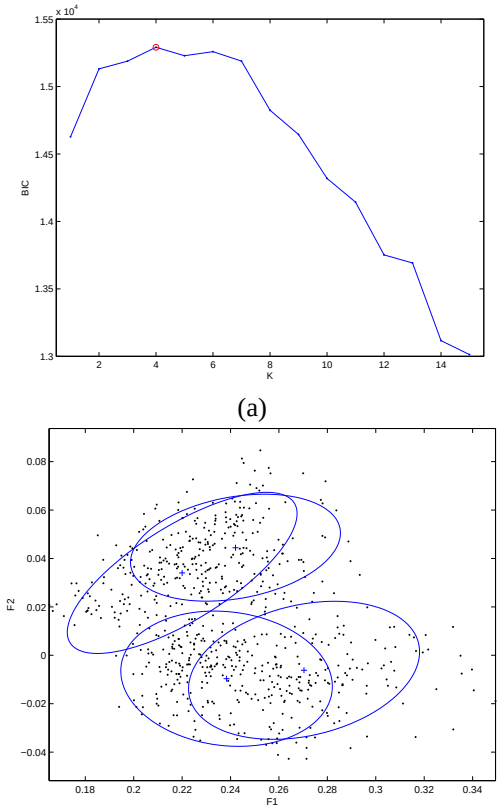


Figure 4: Likelihood model for training examples of walking. (a) BIC values for different K. (b) Mixture model corresponding to the maximum BIC (at $K = 4$).

the various exposures were employed as described in Sect. 4.3. An example MHI for an exposure of 9 frames of a person running is shown in Fig. 3.

To construct the optimal likelihood mixture model for each action class at each level, we randomly partitioned the training features (at each level) into two sets: one for estimating the distributions (using EM) and one for validating its generality (using BIC). We evaluated models having K=1–15 distributions. This process was performed three different times (each time using a random partition). The K-distribution model with the overall largest BIC was selected as the optimal likelihood distribution for the data at that level.

In Fig. 4.a, we show the BIC values as a function of the number of mixture components for the walking training data at level 1. The resulting mixture model corresponding to the maximum BIC (at K=4) is shown in Fig. 4.b for the first two moments. The run and stand classes at this level resulted in 3 and 2 mixture components, respectively.

The hierarchy for the three-class training data terminated

in 15 levels (maximum exposure of .5 sec). The stand class terminated early (as expected) in level 5, and the run class terminated in level 14 (leaving only walk in level 15). In Fig. 5, we show the number of unclassified examples given to each level for each action (failures from the previous level). We show both the actual number of examples employed in each level and the number of deleted examples having no additional frame to make the necessary exposure/MHI for that level. For each action, we see an exponential drop in the number of unclassified examples as we progress down through the levels/exposures. This is the behavior we desire for a rapid recognition system.

As the fraction of recognized class examples at each level is not uniform across the classes, the initial (equal) priors were adapted at each level to reflect the new probabilities of encountering the actions. We show the updated priors at each level (using Eqn. 7) in Fig. 6. Since walk is the only class in the last level (level 15), it is assigned a prior of 1.

5.3. Recognition

In Table 1, we present the recognition results for the testing data in each class, including the average recognition level (where a reliable match was found), the number of examples that found a valid RI match, and the error rate. The percent of testing data for each class that completed and resulted in a valid RI decision ranged from 74–79% (i.e., 21–26% examples could not form the required exposure/MHI and therefore the maximum valid R_{post} was selected). The average level (average exposure time) required by RI to find a reliable action indicator (for the completed actions) was 7 for walk, 6 for run, and 4 for stand. As expected, the stationary stand class required less frames to differentiate itself from walk and run. A histogram showing the distribution of recognition levels for each action is presented in Fig. 7.

The computed error rates were 6% for walk, 16% for run, and 13% for stand (see Table 1). The overall Bayes error for the action context was 12%. These errors were distributed across different levels/exposures (not limited to one short exposure length). The results are encouraging given only very short exposure times to differentiate the actions. In Fig. 8 we show selected starting frames that were correctly labeled (reliable action indicators) at level 1, and other starting frames for longer exposures that were unreliable until reaching their respective action termination levels.

These results were produced from a specifically (yet randomly) selected/partitioned training and testing set at different temporal offsets. To give a more generalized analysis, we generated 10 random split-sample partitions of the entire data set (into training and testing subsets) and averaged the error results (from 10 different sequential RI hierarchies). The average hierarchy depth was 15 levels (min=14, max=17). The average recognition level for walking, running, and standing was 7, 6, and 3. The average error for walking, running, and standing were 8%, 15%, and 10%, and resulted in an average Bayes error of 11% $\pm$.01 (comparable to the previous error).

Action	Examples	Ave. Level	Completed	Error
Walk	708	7	79%	6%
Run	511	6	79%	16%
Stand	258	4	74%	13%

Table 1: Recognition results for testing data.

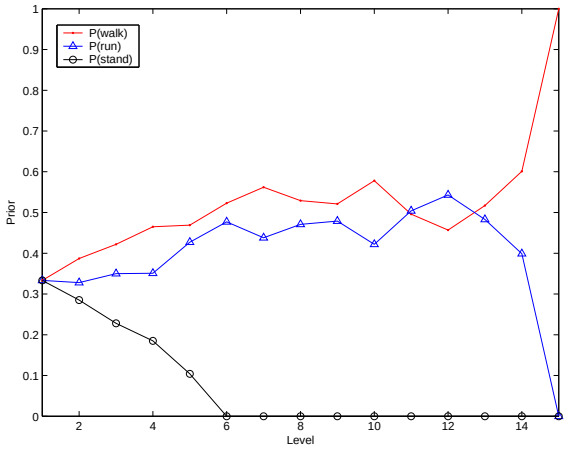


Figure 6: Action priors at each level.

We also compared the R_{post} reliability thresholding method to a maximum-likelihood (ML) approach. In this formulation, we compute the likelihood probability threshold for class $\mathcal{A}_i$ (at a particular level) by computing the negative and positive class thresholds with

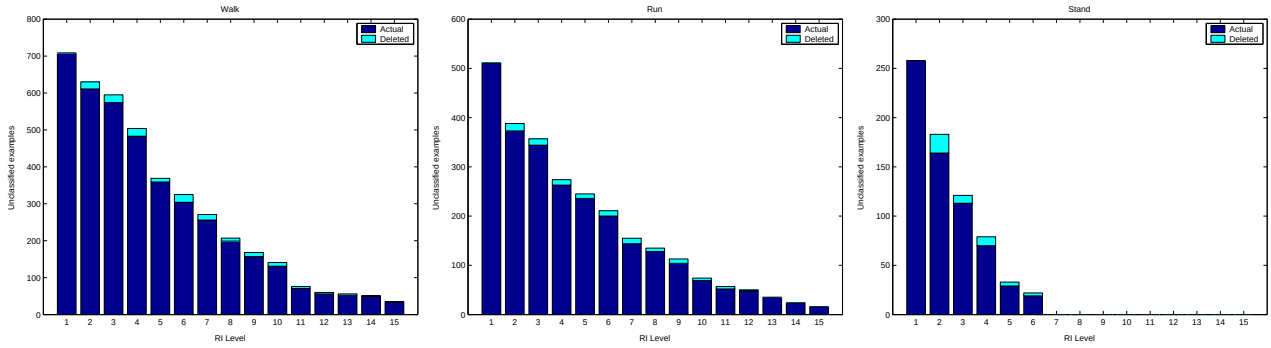
$$T_{neg} = \max p(\mathbf{f}|\mathcal{A}_i, C), \forall \mathbf{f} \in \neg\mathcal{A}_i \quad (14)$$

$$T_{pos} = \min p(\mathbf{f}|\mathcal{A}_i, C) > T_{neg}, \forall \mathbf{f} \in \mathcal{A}_i \quad (15)$$

and set $T_{\mathcal{A}_i} = (T_{neg} + T_{pos})/2$. A feature $\mathbf{f}$ then indicates the presence of action $\mathcal{A}_i$ if $p(\mathbf{f}|\mathcal{A}_i, C) > T_{\mathcal{A}_i}$.

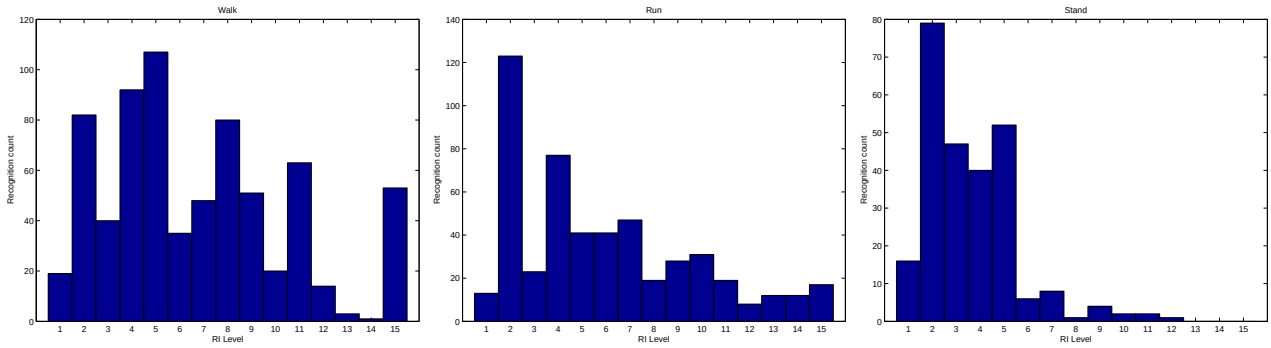
The number of unclassified frames given to each successive level (failures from the previous level) in the ML approach were much higher (worse) than with RI. We show in Fig. 9 the first seven levels for the hierarchies constructed using ML and RI decision thresholds for the walking data. The RI approach gives a much faster reduction of examples as compared to the much slower removal with ML.

We explain the slower ML behavior as the result of ML enforcing a much more strict thresholding policy. When a negative example happens to have a particularly high likelihood for a class, the result is elimination of most of the true positive class examples (having likelihoods smaller than this negative example). In the case of RI, the R_{post} sets



(a) (b) (c)

Figure 5: Number of unclassified examples given to each level for (a) walk, (b) run, and (c) stand. Actual count refers to the number of examples with a valid exposure/MHI at a given level.



(a) (b) (c)

Figure 7: Histogram of recognition levels for (a) walk, (b) run, and (c) stand.

the threshold based on the *ratio* of class posteriors. Even if likelihoods for positive class examples are lower, they may still have higher R_{post} values (much more likely to this class than to the union of the remaining classes).

6. Summary and Future Work

In this paper, we presented a probabilistic framework for rapid-and-reliable action detection. The approach determines the minimum video exposure duration for classification by sequentially lengthening the exposure until it can reliably indicate a particular action. Reliability of a given exposure for discriminating different actions is evaluated using a series of posterior class ratios, where a highly reliable feature (of the exposure) has a high posterior to one class and a low posterior to the remaining classes. The reliability thresholds for each class are automatically learned from negative training examples.

Experimental results demonstrated the efficiency of the approach in terms of reducing the number of training examples used in each successive level. A maximal exposure time of .5 seconds was used to recognize people walking, running, and standing at different styles/offsets/viewpoints, and resulted in a Bayes error of 11%. The R_{post} method of setting class reliability thresholds was also compared to an ML approach, which showed that the RI method more quickly reduced (correctly classified) the training data at each level/exposure.

In future work, we plan to expand the database to include more common actions (particularly those related to surveillance). Additionally, we will extend the framework to use other non-context negative examples and incorporate a likelihood validation at each level to rule out context-violating input and noisy examples early in the search.

References

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

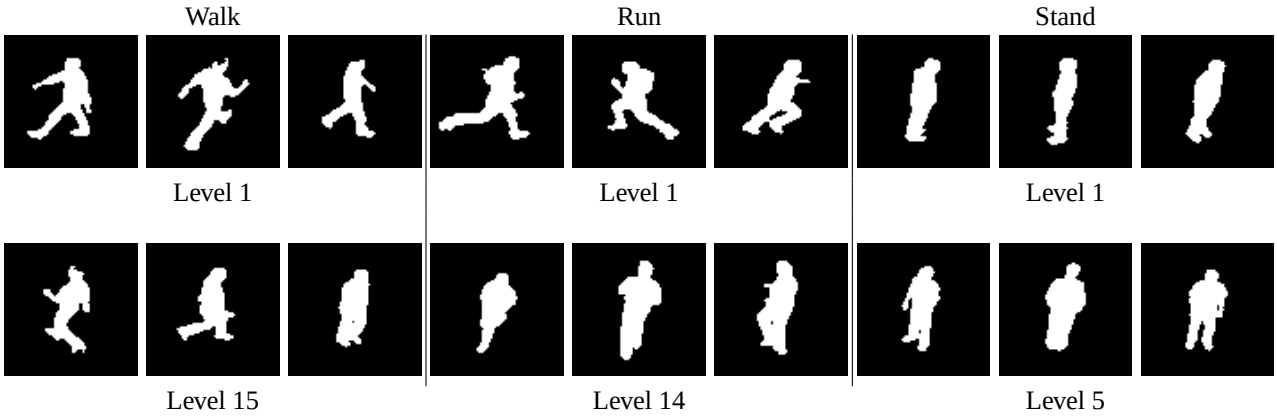


Figure 8: Starting frames for exposures recognized at early and later levels. Top row: starting frames recognized at Level 1. Bottom row: starting frames for exposures recognized at their respective termination levels.

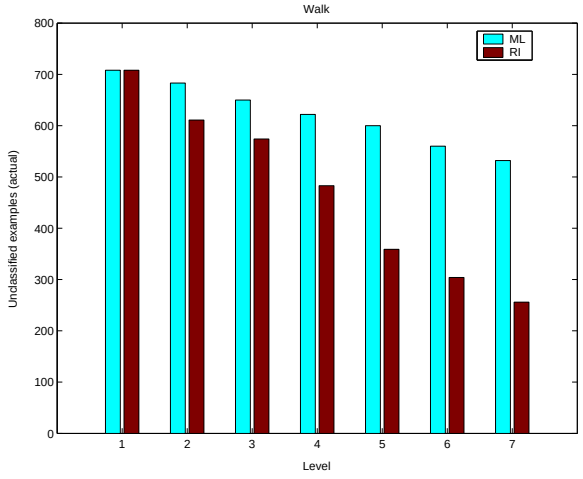


Figure 9: Unclassified ML and RI examples.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

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

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

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

[12] G. Johansson. Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.*, 38:379–393, 1976.

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

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

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

[16] Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frieze patterns. In *Proc. European Conf. Comp. Vis.*, pages 657–671, 2002.

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

[18] 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.

[19] 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.

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

[21] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. Int. Conf. Comp. Vis.*, pages 734–741, 2003.

[22] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.