# Knowing What You Don't Know
## *Novelty Detection for Action Recognition in Personal Robots*

Thomas Moerland, Aswin Chandarr, Maja Rudinac and Pieter Jonker
*Vision-based Robotics, Technical University Delft, Mekelweg 2, Delft, The Netherlands*

Keywords: Action Recognition, Novelty Detection, Anomaly Detection, Computer Vision, Personal Robots.

Abstract: Novelty detection is essential for personal robots to continuously learn and adapt in open environments. This paper specifically studies novelty detection in the context of action recognition. To detect unknown (novel) human action sequences we propose a new method called *background models*, which is applicable to any generative classifier. Our closed-set action recognition system consists of a new skeleton-based feature combined with a Hidden Markov Model (HMM)-based generative classifier, which has shown good earlier results in action recognition. Subsequently, novelty detection is approached from both a posterior likelihood and hypothesis testing view, which is unified as *background models*. We investigate a diverse set of background models: sum over competing models, filler models, flat models, anti-models, and some reweighted combinations. Our standard recognition system has an inter-subject recognition accuracy of 96% on the Microsoft Research Action 3D dataset. Moreover, the novelty detection module combining anti-models with flat models has 78% accuracy in novelty detection, while maintaining 78% standard recognition accuracy as well. Our methodology can increase robustness of any current HMM-based action recognition system against open environments, and is a first step towards an incrementally learning system.

## 1 INTRODUCTION

Recognizing human actions is a very important aspect of robot perception. This becomes even more relevant for personal robots working together with humans in the near future. It is difficult for the robot to learn all different human actions together and have robust recognition performance. Additionally, the subset of actions each robot will need to recognize differs based on the operating environment. Hence an adaptively learning system is necessary, where the robot continuously extends its knowledge about various actions. This process is essential for long term autonomy of personal robots.

In many ways, an action recognition system can be paralleled with speech recognition, with key poses and its sequence similar to alphabets and words. Hence, we borrow motivation from the development of linguistic knowledge in children and translate certain concepts from speech processing into action recognition. It has been studied in psycholinguistics that bootstrapping allows for expansion of cognitive development starting around three years into child growth (Pinker, 1984). Indeed, one of the main components of human intelligence is our adaptivity: we can not only detect what we know, but also identify what we do not know. Moreover, humans use this new input to extend their knowledge, by closing the learning loop (figure 1). In this context, bootstrapping involves equipping the robot with a basis structure and some starting knowledge, from which the robot can detect novel classes and subsequently learn them. Following (Masud et al., 2013), this entire learning process can be modularized into three steps (figure 1):

(i) Anomaly detection (separation): separating videos belonging to known classes from those belonging to unknown classes.

(ii) Cohesion detection: identifying overlapping patterns among buffered anomalous videos identified in (i).

(iii) Retraining: efficiently retraining the ordinary classifier with the new action class, using the detected example videos from (ii).

In this paper, we focus on the first step and investigate new methods for detecting unknown (anomalous) sequences for action recognition systems.

The recent advent of stable 3D imaging technology has strongly increased data quality in action recognition. A landmark paper for 3D action recognition is by (Li et al., 2010) introducing the Microsoft

Moerland, T., Chandarr, A., Rudinac, M. and Jonker, P.
Knowing What You Don't Know - Novelty Detection for Action Recognition in Personal Robots.
DOI: 10.5220/0005677903170327
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 317-327
ISBN: 978-989-758-175-5

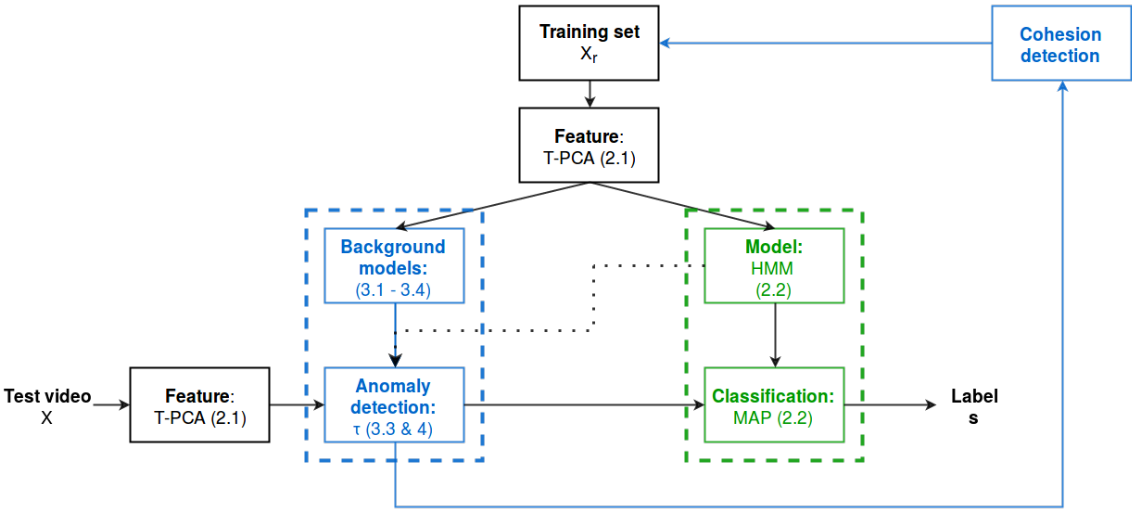

Figure 1: Overall system structure. Proposed novelty detection is shown in blue, all standard action recognition components are shown in black and green. Starting from the training set, we first construct a new compact frame-wise feature based on a Torso-PCA (T-PCA) framework (section 2.1). Then, we train a set of Hidden Markov Models (HMM) with shared keypostures (section 2.2). Each incoming test video is decoded under all class HMM's and assigned according to the maximum-a-posteriori (MAP) rule. We extend this system with a novelty detection module (blue). First, we learn background models from the training set (section 3.1-3.4). These background models are combined with the normal HMM's to obtain a test statistic (*the background corrected likelihood*). In the anomaly detection part we determine a single optimal threshold τ on this test statistic (section 3.3 & 4). When the threshold is exceeded we proceed with standard classification (see figure 3). Else, we identify the video as 'novel/unknown' and buffer it. A cohesion detection module can identify overlap among the buffer videos. When a human supervisor labels the unknown class, we can extend the training set with the new action and close the adaptive learning loop. This paper focusses on the first step of novelty detection: anomaly detection through background models (blue dotted box).

Research Action 3D dataset (MSRA 3D). The authors sample a frame-wise feature from the depth map and use a Hidden Markov Model (HMM) as back-end classifier. Their method has good average recognition accuracy on an inter-subject recognition task (92.9%), but is strongly view-dependent and computationally heavy.

More recently, Shotton et al. introduced the stable extraction of human skeletons from single depth images (Shotton et al., 2013). While the labeling of body parts had been an active research field for many years (Weinland et al., 2011), the direct availability of skeletons raised much interest in the research community. Several papers studied view-invariant skeleton features (Aggarwal and Xia, 2014). Some examples are pairwise joint distances and joint motions (Yang and Tian, 2014) or joint angles and joint angle velocities (Nowozin and Shotton, 2012). An interesting approach combining skeleton and depth map information is called Space-Time Occupancy Patterns (STOP) (Vieira et al., 2012). The authors use the wireframe skeleton to reorientate the depth map to make the subject camera-facing. The feature is constructed from the depth-map occupation over a regular space-time grid, while the back-end classifier is based on a HMM again. Their method still holds the state-of-the-art result on the MSRA 3D inter-subject recognition task (97.5%).

Although the mentioned approaches have made important advancements, they exclusively study their methodology on a closed-set recognition task. Thereby, the system's performance is evaluated on action classes which were also available in the training set. None of the methods consider the occurrence of *unknown* action classes. This specific problem is studied in the machine learning field of *novelty detection*, which is for example reviewed in (Markou and Singh, 2003). While a standard classifier assigns each new instance to the best-fitting class (which is by definition wrong if the video truly belongs to an unknown class), a novelty detection module first tries to identify such novel instances (i.e. anomaly detection).

Anomaly detection has been studied for human activity data before, specifically in the context of abnormal event detection in surveillance video's (as for example reviewed in (Popoola and Wang, 2012)). However, these methods only study the one-class-classification problem between normal and abnormal, for example identifying bikers, skaters or cars on a pedestrian walkway. However, they usually do not try to identify the particular action, i.e. the correct class within the known/normal videos. Thereby they do not need a specific action recognition model, nor could their knowledge system be extended through novelty detection. The current work combines anomaly detection and standard action recognition in one system. In particular, we propose *background models* as an anomaly detection extension to any existing generative classifier like a Hidden Markov Model (HMM)-based recognizer.

Our proposed system structure is shown in figure 1. In the next section we introduce the standard action recognition system based on a compact and view-invariant representation of the human pose from the Kinect's skeletal joint information (2.1) and a HMM-based back-end classifier (2.2) (green box in figure 1).Then, we introduce two dominant views on anomaly detection from speech recognition: posterior probability (3.1) and hypothesis-testing (3.2). These approaches are subsequently unified as background models (3.3). Since this topic has not been studied before for action recognition, we investigate several background models: sum over competing models, filler models, flat models, anti-models and some reweighted combinations of them (3.4). Section 3 thereby covers the blue box in figure 1. The remaining sections of this work present the experimental setup and dataset (section 4), our results including both standard recognition accuracies and various novelty detection results (section 5) and a discussion of our results (section 6).

## 2 ACTION RECOGNITION SYSTEM

In order to investigate novelty detection, we first need a functioning standard recognition system. This consists of two modules; a feature vector which encodes the pose of a given frame into a compact representation, and a generative classifier which uses the encoded features to obtain the probabilities over the trained classes. Our proposed model uses a novel and compact feature vector based on the skeleton information (2.1). The back-end classifier is based on a set of Hidden Markov Models with shared keypostures (2.2).

### 2.1 Representation: Torso-PCA Framework

A good feature is ideally both compact and information-rich. Compact features are especially important for HMM-based back-end classifiers, since it is difficult for these generative probabilistic models to separate signal from possible feature noise. Earlier work on human perception of biological motion has shown that humans can recognize actions by looking only at movements of lights attached to the major joints (Johansson, 1973), implying that tracking of human skeletal poses can provide sufficient information for action recognition.

The availability of real-time skeletal tracking from depth images introduced by (Shotton et al., 2013) has advanced research in action recognition based on this information. The raw skeleton sequence contains the 3D locations of 20 joints at each frame. Many approaches in literature (Yu et al., 2014), (Wang et al., 2014) obtain a feature vector using all pairwise joint distances, velocities and angles. For example, all pairwise joint distances result in a large feature vector (P=190), which not only contains redundant information, but also make the training process difficult due to the dimensionality. Many approaches in literature (Yang and Tian, 2014), (Vieira et al., 2012) employ some dimension reduction technique (usually PCA) to reduce the feature vector length. However, PCA techniques might harm novelty detection, so we construct a novel and compact frame-wise feature based on earlier work by (Raptis et al., 2011).

The raw skeleton sequence contains the 3D locations of 20 joints for each frame (P=60). We construct a more compact frame-wise feature vector (P=30) as illustrated in figure 2. First, we translate the full skeleton to have its origin at the mean of the seven torso joints. Subsequently, we apply PCA on the seven torso joint locations (which form a $7 \times 3$ matrix) to estimate a local coordinate frame with respect to the subject (the three principal axis correspond to the vertical, horizontal and frontal body axis, respectively).

The final feature vector is constructed from the 3D locations of the head, elbows, wrists, knees and ankles augmented with the three rotation angles (yaw, pitch, roll) related to the torso coordinate system. This is based on the assumption that body pose information relevant to actions is majorly encoded in the extremities (ignoring the noisy hand extraction), while the almost rigid body torso can be fully represented by its orientation in 3D space.

Reorientating the full skeleton to make the subject camera-facing has also been implemented. However, this reorientated feature slightly decreased our model performance. This can be explained from the large proportion of camera-facing subjects in our dataset. We therefore choose to use the unrotated feature vectors. Finally, we will report results on the full length feature (P=30) and a PCA-reduced variant (P=10), comparing both when applicable.

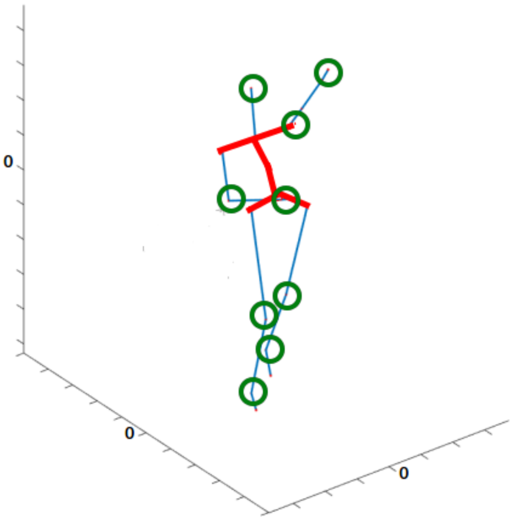

Figure 2: Skeleton-based features. Wireframe skeleton of 20 joints (in blue) is extracted from the Kinect camera following a predescendant method of (Shotton et al., 2013). Full skeleton is translated to have origin at the mean of the seven torso joints (marked by the red lines). Subsequently, a PCA on the torso joint coordinates generates a local coordinate system with respect to the subject's body. The feature vector consists of the 3D coordinates of 9 extremity joints (elbows, wrists, knees, ankles, head; marked by green circles) and the three rotation angles (yaw, pitch, roll) of the torso (as derived from the local coordinate system w.r.t. the world coordinate system).

## 2.2 Classification: HMM with Shared Key Postures

We use the sequences of compact feature vectors obtained from the skeletal data to perform action recognition using a generative model. Hidden Markov Models have shown their major success in speech recognition applications (Gales and Young, 2008), and are now also frequently used as classifiers in action recognition (Weinland et al., 2011). These state-space models naturally handle variation in the speed of performed action. Furthermore, their probabilistic nature allows for novelty detection in low density area's, which will be further pursued in the next section.

We adopt a Hidden Markov Model system with shared key postures between action classes, earlier introduced as an Action Graph (Li et al., 2008). The pooled estimation of the emission model (associated with each key posture) increases model robustness, and furthermore ties the class models together.

The formal definition is as follows: We observe a set of videos $X_r$, $r = 1,2,..N$, with associated class label $s_r \in Q = \{q_1,q_2,q_3,..,q_m\}$, for $m$ different action classes. Each $X_r$, of length $T_r$, has at timepoint $t$ an observation vector $x_{rt}$ of length $P$. Let $W = \{w_1,w_2,..,w_K\}$ denote a set of key postures. In the HMM we assume each feature vector $x_{rt}$ has an associated hidden state variable $z_{rt} \in W$, and the transitions between subsequent hidden states follows a first-order Markov property. Thereby, the transitions between states can be represented as a $K \times K$ transition matrix, where each entry denotes the transition probability between states at subsequent timesteps, i.e. $A_{ij} = P(x_t = w_j | x_{t-1} = w_i)$. We assume K=50 for this work, which is close to the number reported for this dataset elsewhere (Vieira et al., 2012). Furthermore, the relation between the hidden nodes and observation vectors ($P(X|Z = w_k)$), i.e. the emission model, is modeled as a Gaussian with mean vector $\mu_k$ and diagonal covariance matrix $\Sigma_k$.

For each action class $s$ we estimate a separate HMM. However, it is reasonable to assume the key postures and associated emission models are similar between actions. We therefore jointly estimate these parts of the HMM's over the different classes, effectively pooling their contributions. The action classes are discriminated by the class-specific transition matrix $A_s$. The full set of HMM's is thereby defined by the tuple $\Lambda = \{\mu,\Sigma,A\}$, where $A = \{A_1,A_2..A_m\}$.

Under these model assumptions we can write the full data log-likelihood as:

$$\mathcal{L}(X,Z,S|\Lambda) = \sum_{r=1}^{N} \left( \sum_{t=1}^{t_r} \ell(x_{rt}|z_{rt},\mu,\Sigma) + \sum_{t=2}^{t_r} \ell(z_{rt}|z_{r(t-1)},S,A) \right) \tag{1}$$

where $\ell$ denotes a log-probability, and $X$, $S$ and $Z$ denote the video, class and hidden state random variables, respectively. Since the hidden states $Z$ are unobserved, the model is estimated through the well-known Expectation-Maximization (EM) algorithm.

For comparison, we also include a clustered estimation approach. Here, we pool all frame-wise observations vectors in the training set and subsequently cluster these through k-means. Then, we consider each cluster as a key posture, estimating the observation model and transition matrices from their assigned feature vectors. Effectively, we now employ a 'hard' hidden node assignment, compared to the 'soft' hidden node assignment estimated in the EM algorithm.

To perform inference on an incoming video of the test set, we use the maximum-a-posterior (MAP) decision rule:

$$\begin{aligned} \hat{s} = \quad & \text{argmax}_{S \in Q} P(S|X) \\ & \text{argmax}_{S \in Q} \frac{P(X|S)P(S)}{P(X)} \end{aligned} \tag{2}$$

Ordinary speech recognition systems usually assume $P(S)$ is uniform (i.e. no prior on the action class) and ignore $P(X)$, since it does not depend on $S$. Thereby, classification effectively boils down to selecting the class with the highest raw probabilities, $P(X|S)$. These raw likelihoods are obtained through Viterbi decoding.

# 3 NOVELTY DETECTION

The introduced HMM system can only estimate the probability of the input video over the trained classes ($Q$). In this section we introduce novelty detection methodology that has been used in speech recognition and explain our proposed method for novel action detection.

Novelty detection for HMM’s has been previously studied in speech recognition under the name of *Confidence Measures* (CM $\in [0,1]$) (Jiang, 2005). Confidence measures were introduced to post-evaluate the reliability of a recognition decision (as in Equation 2). Since a misrecognition might well be due to a currently unknown class, the goals of CM research and novelty detection are highly overlapping.

We will first introduce the two dominant streams in CM research: posterior probabilities (3.1) and hypothesis testing (3.2). Then we unify both approaches as *background models* (3.3). Since we are the first to investigate novelty detection for action recognition, we will investigate a diverse set of background model types (3.4).

## 3.1 Posterior Probability

A simple and direct way to detect novel classes is to threshold the raw probability $P(X|S)$, as used for assignment in the MAP rule (Equation 2). But this method does not provide a good measure of novelty as $P(X|S)$ is only a relative measure of fit. We do know which class is most likely, but we do not know how good the match really is. In contrary, an absolute and very intuitive measure of fit is the posterior probability of the class given the video: $P(S|X)$. In accordance to Equation 2, we need the marginal probability of the video ($P(X)$) as a normalizing constant. This marginal video probability can be expressed as:

$$P(X) = \sum_{G} P(X|G)P(G) \tag{3}$$

where $G$ denotes the full model space, including the models for all unknown classes. For example, the marginal probability could separate a novel class (high $P(X)$) from a noisy extraction (low $P(X)$). However, the distribution in the unseen model space is not known, and we will need methodology to approximate it. This will be elaborated shortly.

## 3.2 Hypothesis Testing

Another approach to confidence measures for HMM’s was developed independently at AT&T Bell Labs (Sukkar et al., 1996) (Rahim et al., 1997) (Rose et al., 1995). Their work on *utterance verification* casts the problem as a statistical hypothesis test:

$H_0$: $X_r$ is known and correctly recognized
$H_1$: $X_r$ is novel and/or incorrectly recognized

A well-known choice, based on the Neyman-Pearson lemma, is to use the likelihood ratio test (LRT) statistic for testing:

$$LRT = \frac{P(X|H_0)}{P(X|H_1)} \geq \tau \tag{4}$$

As noted by (Jiang, 2005), the major difficulty lies in modelling $H_1$, which is a very composite event with unknown data distribution.

## 3.3 Background Models

We propose both posterior probability and hypothesis testing approaches can be cast in the same framework as *background models*. Both $P(X)$ and $P(X|H_1)$ can be understood as the likelihood of the video under the (partially unobserved) background of the model space. On the log-scale, both methods technically reduce to subtracting the raw likelihood, $P(X|S)$, by a correction factor:

$$\begin{aligned} \ell^{corrected}(X|S) = &\quad \ell(X|S) - \ell(X) \\ = &\quad \ell(X|H_0) - \ell(X|H_1) \end{aligned} \tag{5}$$

As a confirmation of this similarity, both posterior probability and hypothesis testing approaches have independently developed ’filler’ models, by (Kampari and Hazen, 2000) and (Rahim et al., 1997) respectively.

The corrected posterior log-likelihood will be used as the test statistic for anomaly detection. We will identify the video as ’novel’ when the statistic is below a critical threshold $\tau$, i.e. when:

$$\ell(X|S) - \ell(X|M) \leq \tau \tag{6}$$

where M denotes the background model type. If this statistic is higher than $\tau$, we continue with standard class assignment through the MAP decision rule (equation 2). The full test flow is depicted in figure

3. In the next section we introduce different types of background models. Estimation of $\tau$ is discussed in section 4.

### 3.4 Background Model Types

Since we are the first to study anomaly detection methods in the context of a standard action recognition system, we will investigate a diverse set of background models: sum of competing classes, filler models, flat models and anti-models. Filler and flat models are very generic, modelling the distant background of the model space. On the other hand, anti-models approach the closer surroundings of each class. Therefore we also investigate a reweighted combination of them, to combine their advantages.

Background models are themselves Hidden Markov Models, estimated on the same training set as the standard models. However, key postures and emission models obtained in the standard model estimation remain fixed now. All background models except the 'sum over competing hypothesis model' estimate a (class-specific) transition matrix: $\lambda_{type}^{(s)}$. All background models are estimated through EM.

We will denote the video's probability under the background model as $P(X|M_{type}^{(s)})$. The proposed background models are:

(i) Sum over competing hypothesis: This approach is related to the N-best list approaches in speech recognition, like for example in (Kemp and Schaaf, 1997). However, the number of action classes in action recognition is usually smaller, so we can sum over *all* known competing models:

$$\ell(X|M_{sum}) = \log\left(\sum_s P(X|s)\right) \tag{7}$$

(ii) Filler models: Filler models estimate one general transition matrix on all data, which is intended to approximate all humanly possible movements and possible background noise. Speech recognition variants can be found in (Kamppari and Hazen, 2000) and (Rahim et al., 1997):

$$\ell(X|M_{filler}) = \log\left(P(X|\lambda_{filler})\right) \tag{8}$$

(iii) Flat models: Filler models are very generic models for the background, but they are still data dependent. We also include a uniformly initialized transition matrix with each entry equal to $\frac{1}{K}$:

$$\ell(X|M_{flat}) = \log\left(P(X|\lambda_{flat})\right) \tag{9}$$

(iv) Anti-models: As opposed to the previous background models, anti-models are class-specific. They are estimated on all videos *not* belonging to the specific class $s$ (Rahim et al., 1997). Thereby, they are intended to approximate the surroundings of the class' true density area:

$$\ell(X|M_{anti}^{s}) = \log\left(P(X|\lambda_{anti}^{s})\right) \tag{10}$$

(v) Reweighted combinations: To combine the different strengths of the previous approaches, we also include a reweighted mean of filler/flat models with anti-models:

$$\begin{aligned}\ell(X|M_{C1}^{s}) = \log\Big(&0.5\cdot P(X|\lambda_{filler})\\ &+0.5\cdot P(X|\lambda_{anti}^{s})\Big)\end{aligned} \tag{11}$$

$$\begin{aligned}\ell(X|M_{C2}^{s}) = \log\Big(&0.5\cdot P(X|\lambda_{flat})\\ &+0.5\cdot P(X|\lambda_{anti}^{s})\Big)\end{aligned} \tag{12}$$

We use the estimated background likelihoods $\ell(X|M)$ with the novelty detection statistic in Equation 6 to detect previously unknown action sequences as shown in the pipeline of Figure 3.

## 4 DATASET AND EXPERIMENTAL SETUP

We evaluate our proposed method over the publicly available 'Microsoft Research Action (MSRA) 3D' dataset. It contains segmented videos of 20 dynamic actions performed by 10 subjects for ideally 3 repetitions (N=557). Most literature follows the dataset's original paper (Li et al., 2010), where tests are performed in subsets of 8 actions. Since we want to investigate novelty detection, we decide to pool 15 action classes together. These are: horizontal arm wave, hammer, high throw, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw. We only retained videos with three repetitions per subject and action, and also removed some very noisy videos (N=366). For evaluation purposes we use 2/3 of the dataset for training (i.e two of the three videos per subject per action), which corresponds to 'Test 2' of the original paper. Standard recognition results are obtained over three epochs of a 3-fold cross-validation.

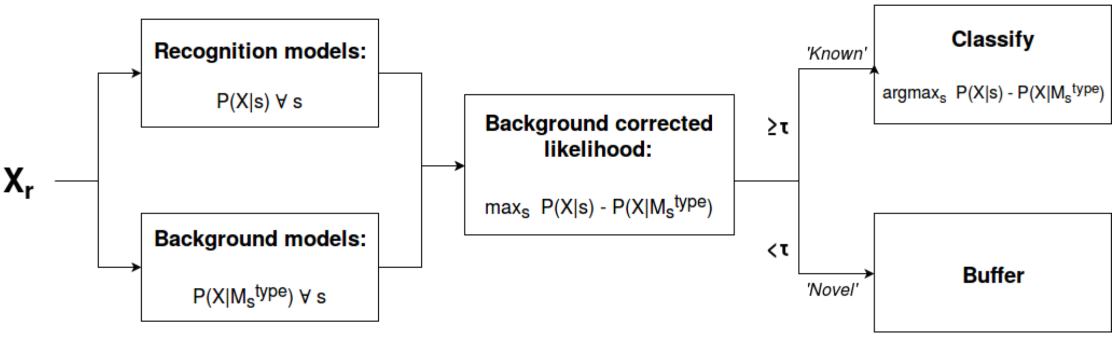

Figure 3: Flow diagram of background model correction and anomaly detection (i.e. expansion of the blue box in figure 1). All probabilities denote their log-scale equivalents. A test video is decoded under standard class models and all background models (possibly class specific). Then, the latter is subtracted from the former to give the background corrected posterior likelihood. The class with the highest posterior likelihood is considered for assignment. When the background corrected posterior likelihood exceeds a threshold τ, we proceed to standard classification through the MAP assignment rule (Equation 2). Else, we refrain from classifying and store the video in a buffer for future processing. Optimization of τ is illustrated in figure 4.

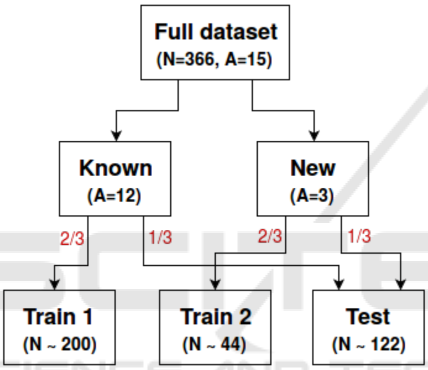

Figure 4: Novelty detection setup. For each run, 3 videos are randomly split off as ‘novel’. Then, in a nested 3-fold cross-validation, HMM’s and background models are trained on known videos (Train 1) and optimal threshold τ is determined on Train 1 and Train 2. Finally, novelty and recognition accuracy are evaluated on Test. N = number of videos in a set, A = number of action classes.

To evaluate novelty performance we will need a double dataset split, as depicted in figure 4. Novelty results are obtained over two full epochs, which each consist of a 5-fold novelty split with nested 3-fold cross-validation (figure 4).

With the introduction of novelty detection, we can also make errors at two levels. Apart from mistakes in the binary novelty module, we can also correctly identify a video as known, but still assign it to the wrong class. The latter is called a *putative* error. However, we are primarily interested in 1) recognition accuracy (percentage of known videos identified to the correct class, i.e. sensitivity) and 2) novelty accuracy (percentage of novel videos correctly identified as novel, i.e. specificity). Therefore, optimal τ will be determined from the largest sum of both accuracies, thereby ignoring the underlying error types.

Table 1: Standard recognition accuracy (no novelty detection) for clustered and EM estimated models.

| **Estimation method** | **Recognition Accuracy** |
|---|---|
| Clustered | 94% |
| EM | 96% |

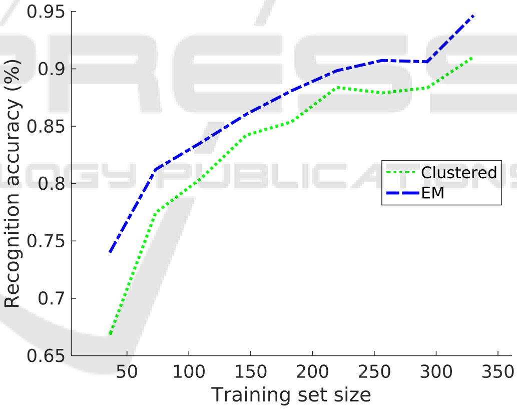

Figure 5: Learning rate: recognition accuracy as a function of the training set size for two estimation methods (EM and clustered). Expectation Maximization is able to learn from the data more quickly. Note that the test results at 2/3 training set size are slightly lower compared to table 1. The setup for this plot is however not solely inter-subject, but makes a random split over the data. Although performance slightly decreases, these results also indicate our method generalizes well for larger training set sizes.

## 5 RESULTS

Recognition accuracy for the standard classification task in shown in table 1. Our novel compact feature (P=10) has accuracy close to the state-of-the-art results on this dataset, although we did use a slightly

Table 2: Overview of recognition accuracy (sensitivity) and novelty accuracy (specificity) for two estimation methods (clustered versus EM) and various background models. Each cell reports accuracy for the PCA-reduced feature vector (P=10) and between brackets the same result for the full-length feature vector (P=30). Optimal performance is obtained for the combined background model (flat + anti-model), with both accuracies at 78%.

| | **Model** | | | |
|---|---|---|---|---|
| | *Clustered* | | *EM* | |
| **Background model** | Sensitivity | Specificity | Sensitivity | Specificity |
| *Raw (none)* | 0.72 (0.73) | 0.60 (0.61) | 0.71 (0.70) | 0.57 (0.63) |
| *Sum* | 0.64 (0.54) | 0.77 (0.77) | 0.66 (0.59) | 0.74 (0.73) |
| *Filler* | 0.73 (0.73) | 0.66 (0.67) | 0.73 (0.75) | 0.58 (0.52) |
| *Flat* | 0.68 (0.66) | 0.77 (0.78) | 0.71 (0.73) | 0.74 (0.69) |
| *Anti-model* | 0.77 (0.73) | 0.77 (0.70) | 0.73 (0.75) | 0.69 (0.62) |
| *Combination 1 (filler + anti)* | 0.76 (0.72) | 0.75 (0.71) | 0.73 (0.77) | 0.68 (0.62) |
| *Combination 2 (flat + anti)* | **0.78** (0.75) | **0.78** (0.75) | 0.73 (0.77) | 0.73 (0.65) |

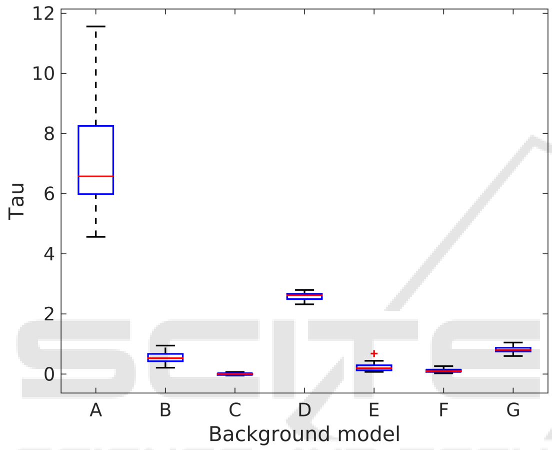

Figure 6: Consistency of τ. Boxplots show the distribution of τ over multiple splits for the seven background models (A-G) as reported in the rows of table 2, respectively. Results are obtained over 2 full epochs (30 runs, see figure 4) on the clustered model with P=10. On each run, we determine a single value of τ. Clearly, the raw likelihood (model A) has trouble generalizing over different data splits. However, all background models (B-G) improve generalization, since the variance in τ decreases.

different test set-up (see section 4). The ability of our estimation methods to learn from smaller amounts of data is shown in figure 5. The graph indicates EM estimation is able to learn from the data more quickly, although both methods eventually approach the same recognition accuracy.

Novelty detection results are reported in table 2. Results do not differ between the PCA-reduced (P=10) and full-length (P=30) feature vectors. The raw posterior likelihood is able to identify around 72% of the known videos in the correct class, while also detecting 60% of the novel videos. Obviously, recognition accuracy has decreased compared to table 1, since we augmented the problem by adding a set of videos from classes unavailable during training.

Table 3: Optimal novelty detection results for the clustered standard model with flat & anti-model background (bold result in table 2). The results illustrate that we hardly make any putative errors (i.e. known videos assigned to the wrong class). The decrease in recognition accuracy from about 95% (table 1) to 78% can be almost fully attributed to misrecognized novel videos.

| | True label | |
|---|---|---|
| Assigned label | **Known** | **Novel** |
| **Known (correct)** | 78% | 22% |
| **Known (wrong)** | 1% | - |
| **Novel** | 21% | 78% |

The different background models all improve performance, but in different ways. Optimal performance is achieved for the combined background of flat and anti-model, which reaches novelty and recognition accuracy of both 78%. Interestingly, the clustered estimation seems to outperform EM for novelty detection in general.

Table 3 shows the underlying errors of our optimal novelty detection result. Although we optimized τ to maximize the sum of recognition accuracy and novelty accuracy, we can observe a clear difference in the type of errors our system makes. Putative errors, i.e. known videos assigned to the wrong class, occur only for 1% of the known videos. Thereby, the recognition accuracy considering only the known classes (i.e. a closed-set recognition problem) has actually increased compared to table 1. We could have expected this result, since our inspiration (confidence measures in speech recognition) was actually developed to identify misrecognitions.

A closer inspection of the scaling of τ is provided in figure 7. The ROC-curve (top) shows both accuracies for different values of τ. The bottom plot shows the distribution of the background corrected likelihood for known and new classes. As expected, the distribution under the known class has a higher mean posterior likelihood. We see an overlapping area, cor-

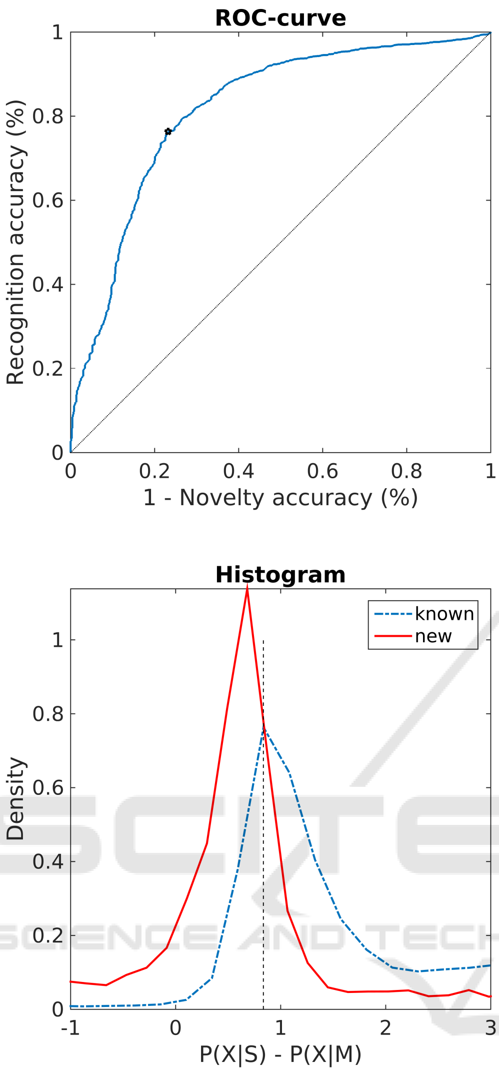

Figure 7: Scaling of $\tau$. Top: ROC-curve showing recognition and novelty accuracy for various levels of $\tau$. Model is cluster estimated with combined (flat + anti-model) background correction. Optimal joint performance is marked with an asterix (the associated value of $\tau$ is visible in the bottom plot). Bottom: Distribution of background corrected likelihood for known (blue dashed line) and novel (red solid line) videos. Optimal $\tau$ is indicated by the vertical dashed line.

responding to the 22% errors on both sides.

Finally, we investigate the ability of $\tau$ to generalize over different novelty settings. Figure 6 shows the distribution of the optimal $\tau$ for all background models over various dataset splits (according to figure 4). We see the raw likelihood without background correction (model A) has difficulty generalizing over different splits, i.e. the variance of the optimal $\tau$ is large. On the other hand, all background methods clearly improve the consistency of $\tau$, indicating the background models do systematically correct the different scalings of the raw likelihood.

## 6 DISCUSSION AND CONCLUSION

To our knowledge, our work is the first to report on novelty detection in the context of a standard action recognition system. Our methodology can assign 78% of the known videos to the correct class, while also identifying 78% of the unknown videos as novel. Furthermore, our background model methodology shows consistent results over various dataset splits, indicating the method should generalize well to different settings. Background models can be easily implemented on any HMM-based action recognition system, providing it with robustness against open-set environments.

An interesting aspect of our results is the relatively good performance of the flat background model. This model was the only data-independent background approach. Most literature on novelty detection tries to re-use the dataset in some smart way. The good results of the flat model touch upon a fundamental challenge in novelty detection: ‘you can not model the unknown (new classes) from the known (data)’.

We think the results in table 3 could give a motivation to use our methodology even for closed-set environments. As we mentioned before, the background model system can also be used as a confidence measure to identify misrecognitions. Considering only the closed-set problem (i.e. first column of table 3), we would refuse to classify 21% of the videos, but for the assigned videos we can be very certain that the class is correct.

The decrease in recognition accuracy from 96% for closed-set recognition to 78% for open-set recognition nicely illustrates the inevitable trade-off in novelty detection. As table 3 clearly shows, the bottleneck of this decrease is in the novelty detection module. By including a set of unknown videos, we strongly increased the difficulty of the classification task. However, real-life is by definition an open-set, and any system (like a personal robot) ignoring this problem will see their good closed-set recognition performance strongly decrease in practical application.

The test setup (figure 4) could be further improved. Ideally, one would not use unknown videos from the same class for optimizing $\tau$ and evaluating performance. However, the size of our dataset (15 action classes) did not allow such a ’triple split’, since anomaly detection will by definition need a substantial amount of known classes (i.e. the basic knowledge, being 12 classes in this experiment). We used this dataset (MSRA 3D) since it allowed us to compare our closed-set method with the state-of-the-art in the field. However, we do not expect overfitting was a problem for the scaling of $\tau$. In particular, the consistency of $\tau$ over various dataset splits (figure 6) would be highly unexpected if overfitting was a serious problem.

In conclusion, we identify three purposes for our anomaly detection methodology based on background models: 1) increased accuracy in *closed-set* recognition tasks by acting as a confidence measure, 2) increased robustness against *open-set* problems by filtering of unknown videos and 3) as a first step towards *adaptive* learning by closing the learning loop of figure 1. Due to the large resemblance of human intelligence, novelty detection can significantly extend both robotic functionality and human-robot interaction. We intend to implement the novelty detection methodology on our personal robot (Chandarr et al., 2013) and tackle the challenges posed by unconstrained motions and environments.

## ACKNOWLEDGEMENTS

The authors want to thank Tim van Erven and David Tax for the useful discussions.

## REFERENCES

Aggarwal, J. and Xia, L. (2014). Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, 48:70–80.

Chandarr, A., Bruinink, M., Gaisser, F., Rudinac, M., and Jonker, P. (2013). Towards bringing service robots to households: Robby ,Lea smart affordable interactive robots. In *IEEE/RSJ International Conference on Advanced Robotics (ICAR 2013)*.

Gales, M. and Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception &amp; psychophysics*, 14(2):201–211.

Kamppari, S. O. and Hazen, T. J. (2000). Word and phone level acoustic confidence scoring. In *ICASSP*, pages 1799–1802. IEEE.

Kemp, T. and Schaaf, T. (1997). Estimating confidence using word lattices. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *EUROSPEECH*. ISCA.

Li, W., Zhang, Z., and Liu, Z. (2008). Expandable data-driven graphical modeling of human actions based on salient postures. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1499–1510.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3D points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14.

Markou, M. and Singh, S. (2003). Novelty detection: a review–part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

Masud, M. M., Chen, Q., Khan, L., Aggarwal, C. C., Gao, J., Han, J., Srivastava, A. N., and Oza, N. C. (2013). Classification and adaptive novel class detection of feature-evolving data streams. *IEEE Trans. Knowl. Data Eng.*, 25(7):1484–1497.

Nowozin, S. and Shotton, J. (2012). Action points: A representation for low-latency online human action recognition. *Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68*.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Popoola, O. P. and Wang, K. (2012). Video-based abnormal human behavior recognition—A review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):865–878.

Rahim, M. G., Lee, C.-H., and Juang, B.-H. (1997). Discriminative utterance verification for connected digits recognition. *Speech and Audio Processing, IEEE Transactions on*, 5(3):266–277.

Raptis, M., Kirovski, D., and Hoppe, H. (2011). Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '11, pages 147–156, New York, NY, USA. ACM.

Rose, R. C., Juang, B.-H., and Lee, C.-H. (1995). A training procedure for verifying string hypotheses in continuous speech recognition. In *ICASSP*, pages 281–284. IEEE Computer Society.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Sukkar, R., Lee, C.-H., et al. (1996). Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 4(6):420–429.

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., and Campos, M. F. (2012). STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 252–259. Springer.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2014). Learning actionlet ensemble for 3D human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(5):914–927.

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241.

Yang, X. and Tian, Y. (2014). Effective 3D action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2–11.

Yu, G., Liu, Z., and Yuan, J. (2014). Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*.