Informed Democracy: Voting-based Novelty Detection for Action Recognition

Alina Roitberg*
alina.roitberg@kit.edu
Ziad Al-Halah*
ziad.al-halah@kit.edu
Rainer Stiefelhagen
rainer.stiefelhagen@kit.edu

Karlsruhe Institute of Technology,
76131 Karlsruhe,
Germany

Abstract

Novelty detection is crucial for real-life applications. While it is common in activity recognition to assume a closed-set setting, *i.e.* test samples are always of training categories, this assumption is impractical in a real-world scenario. Test samples can be of various categories including those never seen before during training. Thus, being able to know *what we know* and *what we don't know* is decisive for the model to avoid what can be catastrophic consequences. We present in this work a novel approach for identifying samples of activity classes that are not previously seen by the classifier. Our model employs a voting-based scheme that leverages the estimated uncertainty of the individual classifiers in their predictions to measure the novelty of a new input sample. Furthermore, the voting is privileged to a subset of *informed* classifiers that can best estimate whether a sample is novel or not when it is classified to a certain known category. In a thorough evaluation on UCF-101 and HMDB-51, we show that our model consistently outperforms state-of-the-art in novelty detection. Additionally, by combining our model with off-the-shelf zero-shot learning (ZSL) approaches, our model leads to a significant improvement in action classification accuracy for the generalized ZSL setting.

1 Introduction

Human activity recognition from video is a very active research field, with a long list of potential application domains, ranging from autonomous driving to security surveillance [1, 25]. However, the vast majority of published approaches are developed under the assumption that all categories are known a priori [5, 11, 13, 34, 40, 41]. This *closed set* constraint represents a significant bottleneck in the real world, where the system will probably encounter samples from various categories including those never seen during development. The set of possible actions is dynamic by its nature, possibly changing over time. Hence, collecting and maintaining large scale application-specific datasets of video data is especially costly and impractical. This raises a crucial need for the developed models to be able to identify cases where they are faced with samples out of their knowledge domain. In this work, we

* Equal contribution

explore the field of activity recognition under *open set* conditions [2, 30, 31], a setting which has been little-explored before especially in the action recognition domain [19].

In an open world application scenario, an action recognition model should be able to handle three different tasks: 1) the standard classification of previously seen categories; 2) knowledge transfer for generalization to new unseen classes (e.g. through zero-shot learning); 3) and knowing how to automatically discriminate between those two cases. The third component of an open set model lies in its ability to identify samples from unseen classes (*novelty detection*). This is closely linked to the classifier's confidence in its own predictions, *i.e.* how can we build models, that know, what they do not know? A straight-forward way is to employ the *Softmax* output of a neural network (NN) model as the basis for a rejection threshold [27, 29]. Traditionally, action recognition algorithms focus on maximizing the top-1 performance on a static set of actions. Such optimization leads to *Softmax* scores of the winning class being strongly biased towards very high values [9, 14, 22, 39]. While giving excellent results in closed set classification, such overly self-confident models become a burden under open set conditions. A better way to asses NN's confidence, is to rather predict the probability distribution with Bayesian neural networks (BNN). Recently, Gal *et al.* [9] introduced a way of efficiently approximating BNN modeled as a Gaussian Process [28] and using dropout-based Monte-Carlo sampling (MC-Dropout) [9]. We leverage the findings of [9] and exploit the *predictive uncertainty* in order to identify activities of previously unseen classes.

This work aims at bringing conventional activity recognition to a setting where new categories might occur at any time and has the following main contributions: 1) We present a new model for novelty detection for action recognition based on the predictive uncertainty of the classifiers. Our main idea is to estimate the *novelty* of a new sample based on the uncertainty of a selected group of output classifiers in a voting-like manner. The choice of the voting classifiers depends on how confident they are in relation to the currently predicted class. 2) We adapt zero-shot action recognition models, which are conventionally applied solely on samples of the *unseen* classes, to the generalized case (*i.e.* open set scenario) where a test sample may originate from either known or novel categories. We present a generic framework for generalized zero-shot action recognition, where our novelty detection model serves as a filter to distinguish between seen and novel categories, passing the sample either to a standard classifier or a zero-shot model accordingly. 3) We extend the custom evaluation setup for action recognition to the open-set scenario and formalize the evaluation protocol for the tasks of novelty detection and zero-shot action recognition in the generalized case on two well-established datasets, UCF-101 [36] and HMDB-51 [15]. The evaluation shows, that our model consistently outperforms conventional NNs and other baseline methods in identifying novel activities and was highly successful when applied to generalized zero-shot learning.

2 Related Work

Novelty Detection Various machine learning methods have been used for quantifying the *normality* of a data sample. An overview of the existing approaches is provided by [6, 24]. A lot of today's novelty detection research is handled from the probabilistic point of view [16, 21, 24, 35], modeling the probability density function (PDF) of the training data, with Gaussian Mixture Models (GMM) being a popular choice [24]. The One-class SVM introduced by Schölkopf *et al.* [33] is another widely used unsupervised method for novelty

detection, mapping the training data into the feature space and maximizing the margin of separation from the origin. Anomaly detection with NNs has been addressed several times using encoder-decoder-like architectures and the reconstruction error [43]. A common way for anomaly detection is to threshold the output of the neuron with the highest value [12, 17, 29]. Recently, Hendrycks *et al*. [12] presented a baseline for deep-learning based visual recognition using the top-1 Softmax scores and pointed out, that this area is under-researched in computer vision.

The research of novelty detection in videos has been very limited. A related topic of *anomaly detection* has been studied for very specific applications, such as surveillance [17, 24] or personal robotics[19]. Surveillance however often has anomalies, such as *Robbery* or *Vandalism*, present in the training set in some form [20, 38] which violates our open-set assumption. The work most similar to ours is the one of Moerland *et al*. [19] where Hidden-Markov-Model is used to detect unseen actions from skeleton features. However, [19] considers only a simplified evaluation setting using only a single *unseen* action category in testing. In contrast to [19] our model is based on a deep neural architecture for detecting novel actions which makes it applicable to a wide range of modern action recognition models. Furthermore, we consider a challenging evaluation setting on well-established datasets where novel classes are as diverse as those seen before. Additionally, we go beyond novelty detection and evaluate how well our model generalizes to classifying novel classes through zero-shot learning. Our model leverages approximation of BNN using MC-Dropout as proposed by Gal *et al*. [9], which has been successfully applied in semantic segmentation [14] and active learning [10]. We extend the BNN approximation to the context of open set action recognition where we incorporate the uncertainty of the output neurons in a voting scheme for novelty detection.

Zero-Shot Action Recognition Research on human activity recognition under open set conditions has been sparse so far. A related field of Zero-Shot Learning (ZSL) attempts to classify new actions without any training data by linking visual features and the high-level semantic descriptions of a class, *e.g*. through action labels. The description is often represented with word vectors by a skip-gram model (*e.g*. *word2vec* [18]) previously trained on a large-scale text corpus. ZSL for action recognition gained popularity over the past few years and has also been improving slowly but steadily [26, 42, 45, 46, 47]. In all of these works, the categories used for training and testing are disjoint and the method is evaluated on unfamiliar actions only. This is not a realistic scenario, since it requires the knowledge of whether the activity belongs to a known or novel category a priori. Generalized zero-shot learning (GZSL) has been recently studied for image recognition and a drastic performance drop of classical ZSL approaches such as ConSE [23] and Devise [7], has been reported [44]. As the main application of our novelty detection approach, we implement a framework for ZSL in the generalized case and integrate our novelty detection method to distinguish between known and unknown actions.

3 Novelty Detection via Informed Voting

We present a new approach for novelty detection in action recognition. That is, given a new video sample $\mathbf{x}$, our goal is to find out whether $\mathbf{x}$ is a sample of a previously known category or if it belongs to a novel action category not seen before during training.

Let $A = \{A_1, ...A_K\}$ be the set of all K *known* categories in our dataset. Then $p(A_i|\mathbf{x})$ is the classifier probability of action category A_i given sample $\mathbf{x}$. Conceptually, our novelty

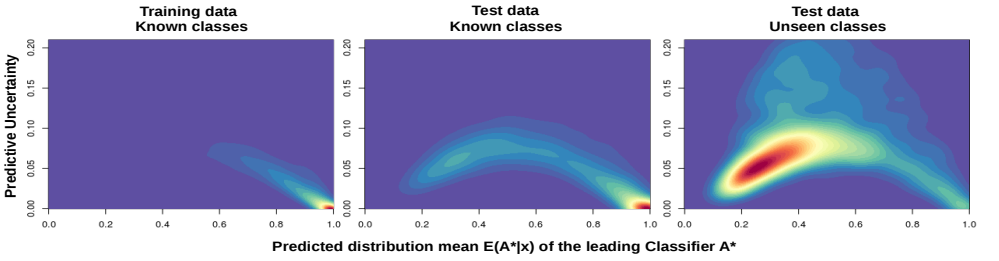

Figure 1: Distribution of predictive mean and uncertainty as a 2-D histogram of the leading classifier (highest predictive mean) for the input with known and unseen actions (HMDB-51 dataset). Red denotes common cases (high frequency), blue denotes unlikely cases.

detection model is composed of two main components: 1) the leader and 2) the council. The leader refers to the classifier with the highest confidence score in predicting the class of a certain sample $\mathbf{x}$. For example, in classification neural networks it is common to select the leader based on the highest softmax prediction score. The leader votes for sample $\mathbf{x}$ being of its own category and assuming that the class of $\mathbf{x}$ is one of the known categories, *i.e.* $\text{class}(\mathbf{x}) = A^* \in A$. The council, on the other hand, is a *subset* of classifiers that will help us validating the decision of a specific leader. In other words, the council members of a leader representing the selected class A^* are a subset of the classifiers representing the rest of the classes, *i.e.* $C_{A^*} \subseteq A \setminus \{A^*\}$. These members are elected for each leader individually, *i.e.* each category classifier in our model has its own council. A council member is selected based on its certainty variance in relation to a leader. Whenever a leader decides on the category of a sample $\mathbf{x}$, its council will convene and vote on the leader decision. Then, the council members will jointly decide whether the leader made the correct decision or it was mistaken because the sample is actually from a novel category.

Next, we explain in details how we measure the uncertainty of a classifier (Section 3.1); choosing a leader and its council members (Section 3.2); and, finally, the novelty voting procedure given new sample (Section 3.3).

3.1 Measuring Classifier Uncertainty

In this section, we tackle the problem of quantifying the *uncertainty* of a classifier given a new sample. The estimated uncertainty is leveraged later by our model to select the council members as we will see in Section 3.2.

In the context of deep learning, it is common to consider the single point estimates for each category, represented by the output of the softmax layer, as a confidence measure [8, 12, 17, 29]. However, this practice has been highlighted in literature to be inaccurate since a model can be highly uncertain even when producing high prediction scores [22, 39]. Bayesian neural networks (BNNs) offer us an alternative to the point estimate models and are known to provide a well calibrated estimation of the network uncertainty in its output. Given the network parameters ω and a training set S, the predictive probability of the BNN is obtained by integrating over the parameter space. The prediction $p(A_i|\mathbf{x},S)$ is therefore the mean over all possible parameter combinations weighted by their posterior probability:

$$p(A_i|\mathbf{x},S) = \int_{\omega} p(A_i|\mathbf{x},\omega)p(\omega|S)d\omega \tag{1}$$

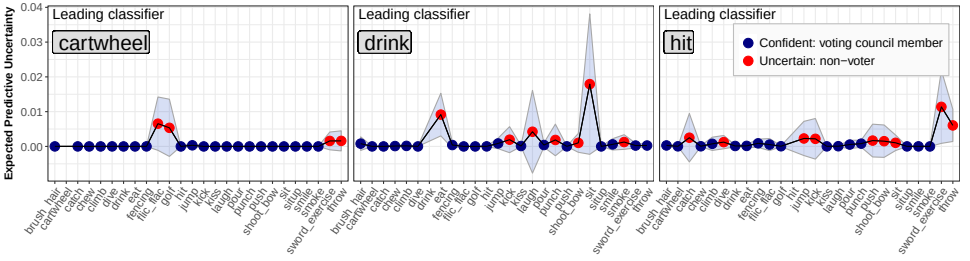

Figure 2: Council members and uncertainty statistics for three different leaders (HMDB-51). The classifier's *average* uncertainty and its *variance* (area surrounding the point) illustrate how it changes its belief in the leader for different data inputs. Blue points are in the council of the current leader, while red points are classifiers that did not pass the credibility threshold.

However, BNNs are known to have a difficult inference scheme and high computation cost [9]. Therefore, we leverage the robust model proposed by [9] to approximate the predictive mean and uncertainty of the BNN posterior distribution with network parameters modeled as a Gaussian Process (GP). This method is based on dropout regularization [37], a widely used technique which has proven to be very effective against overfitting. That is, it leverages the dropout at each layer in the network to draw the weights from a Bernoulli distribution with probability p. At test time, the dropout is iteratively applied for M forward passes for each individual sample. Then, the statistics of the neural network output represents a Monte-Carlo (MC) approximation of the neuron's posterior distribution This approach is referred to as MC-Dropout [9].

Specifically, let $\mathbf{x}$ be a representation generated by a convolutional neural network (CNN) for an input sample z. We add a feedforward network on top of the CNN with two fully-connected layers with weight matrices W_1 and W_2. Instead of using a deterministic *Softmax* estimate in a single forward pass as it is common with CNNs, we now compute the mean over M stochastic iterations as our prediction score:

$$\mathbb{E}(A_i|\mathbf{x}) \approx \frac{1}{M}\sum_{m=1}^{M} \text{softmax}(\text{relu}(\mathbf{x}^T D_1 W_1 + \mathbf{b}_1) D_2 W_2), \tag{2}$$

where $relu(\cdot)$ is the rectified linear unit (ReLU) activation function, $\mathbf{b}_1$ is the bias vector of the firs layer. Additionally, D_1 and D_2 are diagonal matrices where the diagonal elements contain binary values, such that they are set to 1 with probability $1-p$ and otherwise to 0.

We further empirically compute the model's *predictive uncertainty* as the distribution variance:

$$U(A_i|\mathbf{x}) \approx s^2 = \frac{1}{M-1}\sum_{m=1}^{M} [\text{softmax}(\text{relu}(\mathbf{x}^T D_1 W_1 + \mathbf{b}_1) D_2 W_2) - \mathbb{E}(A_i|\mathbf{x})]^2 \tag{3}$$

Fig. 1 shows how predictive mean and uncertainty are distributed for samples of known and novel classes. The plot depicts clearly different patterns for the resulting probability distributions in these two cases which illustrates the potential of Bayesian uncertainty for novelty detection.

3.2 Selecting the Leader and its Council

Now that we can estimate the confidence and uncertainty of each category classifier in our model, we describe in this section how to choose the leader and select it council members.

The Leader. Rather than selecting the leader using a point estimate based on the softmax scores of the output layer, we leverage here the more stable dropout-based estimation of the prediction mean. Hence, the leader is selected as the classifier with highest expected prediction score over M sampling iterations:

$$A^* = \underset{A_k \in A}{\operatorname{argmax}}\ \mathbb{E}(A_i|\mathbf{x}), \tag{4}$$

where $\mathbb{E}(A_i|\mathbf{x})$ is estimated according to Eq. 2.

The Council. The leader by itself can sometimes produce highly confident predictions for samples of unseen categories [39]. Hence, we can not rely solely on the leader confidence to estimate whether a sample is of a novel category or not. Here, the rest of the classifiers can help in checking the validity of the leader's decision. We notice that these classifier exhibit unique patterns in regard to a certain leader. They can be grouped into two main groups: the first shows high uncertainty when the leader is correctly classifying a sample; while the second shows a very low uncertainty and are in agreement with the leader.

Guided by this observation, we select the members of the Council C_A^* for a certain leader A^* based on their uncertainty variance in regards to samples of the leader's category, *i.e.* $\mathbf{x} \in A^*$. In other words, those classifiers that exhibit very low uncertainty when the leader is classifying samples of its own category are elected to join its council. During the training phase, we can select the council members for each classifier in our model. Here, we randomly split the initial set into a training set S_{train} which is used for model optimization and parameter estimation, and a holdout set $S_{holdout}$ which is used for choosing the council member for all the classifiers iteratively. Specifically, we use a 9/1 split for the training and the holdout splits. We first estimate the parameters of our deep model ω using S_{train}. Then, we evaluate our model over all samples from $S_{holdout}$. For each category classifier in our model, we construct a set of true positive samples $S_{tp}^{A_i} \subseteq S_{holdout}$. For each sample $\mathbf{x_n} \in S_{tp}^{A_i}$, we estimate the uncertainty $U(A_j|\mathbf{x_n})$ of the rest of the classifiers $A_j \in A \setminus \{A_i\}$ using the MC-Dropout approach. Then, the variance of these classifiers' uncertainty is estimated as:

$$Var(A_j|A_i) = \frac{1}{N}\sum_{n=1}^{N}(U(A_j|\mathbf{x}_n) - \mathbb{E}[U(A_j|\mathbf{x})])^2 \tag{5}$$

where $N = |S_{tp}^{A_i}|$ and $\mathbb{E}[U(A_j|\mathbf{x})]$ is the expectation of the uncertainty of the classifier A_j over samples $\mathbf{x} \in S_{tp}^{A_i}$. Finally, classifiers with a variance lower than a fixed credibility threshold $Var(A_j|A_i) < c$ are then elected as members of A_i council.

Fig. 2 shows three leaders and their elected councils according to our approach. We see, for example, that eight classifiers did not pass the credibility threshold for the leader *drink* and were excluded from its council. The variance of the uncertainty is especially high for *sit* and *eat* in this case. This is expected since those actions often occur in a similar context.

3.3 Voting for Novelty

Given the trained deep model and the sets of all council members from the previous step, we can now generate a novelty score for a new sample $\mathbf{x}$ as follows. First, we calculate the prediction mean $\mathbb{E}[p(A_i|\mathbf{x})]$ and uncertainty $U(A_i|\mathbf{x})$ of all the action classifiers using M stochastic forward passes and MC-Dropout. Then, the classifier with the maximum predicted mean is chosen as the leader. Finally, the council members of the chosen leader vote for the novelty of sample $\mathbf{x}$ based on their estimated uncertainty (see Algorithm 3.3).

Examples of such voting outcome for three different leaders are illustrated in Fig. 3. In case of category *cartwheel*, we can see that when the leader is voting indeed for the correct category, all council members show low uncertainty values therefore resulting in a low novelty score, as uninformed classifiers (marked in red) are excluded. However, we observe very different measurements for an example from an unseen category *clap* which is also predicted as *cartwheel*. Here, multiple classifiers which are in the council (marked in blue) show unexpected high uncertainty values (e.g. *eat*, *laugh*), therefore discrediting the leader decision and voting for a high novelty score.

Algorithm Novelty Detection by Voting of the Council Neurons

Input: Input sample $\mathbf{x}$, Classification Model ω, K sets of *Council* members for each *Leader*: $\{C_{A^1}, ..., C_{A^K}\}$

Output: Novelty score $\upsilon(\vec{x})$

1: **Inference using MC-Dropout**
Preform M stochastic forward passes: $p^m_{A_i} = p(A_i|\mathbf{x}, \omega^m)$;

2: **for all** $A_i \in A$ **do**

3: Calculate the prediction mean and uncertainty: $\mathbb{E}(p(A_i|\mathbf{x}))$ and $U(A_i|\mathbf{x})$

4: **end for**

5: Find the *Leader*: $A^* = \underset{A_k \in A}{\operatorname{argmax}}\ p(A_i|\mathbf{x})$

6: Select the *Council*: C_{A*}

7: Compute the *novelty score* : $\upsilon(\mathbf{x}) = \frac{\sum_{A_i \in C_{A^*}} U(A_i|\mathbf{x})}{|C_{A^*}|}$

Model variants. We refer to our previous model as the *Informed Democracy* model since voting is restricted to the council members which are chosen in an informed manner to check the decision of the leader. In addition to the previous model, we consider two other variants of our model:

1. The *Uninformed Democracy* model: Here, there is no council and all classifiers have the right to vote for any leader. Hence, step 7 in Algorithm 3.3 is replaced with $\upsilon(\mathbf{x}) = \frac{\sum_{A_i \in A} U(A_i|\mathbf{x})}{K}$.
2. The *Dictator* model: unlike the previous model, this one leverages only the leader's uncertainty in its own decision to predict the novelty of the sample, *i.e.* $\upsilon = U(A^*|\vec{x})$.

Open set and zero-shot learning Once our model generated the novelty score $\upsilon(\mathbf{x})$, we can decide whether $\mathbf{x}$ is a sample from a novel category or not using a sensitivity threshold τ. This threshold can be estimated from a validation set using the equal error rate of the receiver operating characteristic curve (ROC). Then, if $\upsilon(\mathbf{x}) < \tau$ the *Council* votes in favor of the *Leader* and its category is taken as our final classification result. Otherwise, an *unknown* activity class has been identified. In this case, the input could be passed further to a module in charge of handling unfamiliar data, such as a zero-shot learning model or a user to give the sample a new label in the context of active learning.

4 Evaluation

Evaluation setup Since there is no established evaluation procedure available for action recognition in open-set conditions, we adapt existing evaluation protocols for two well-

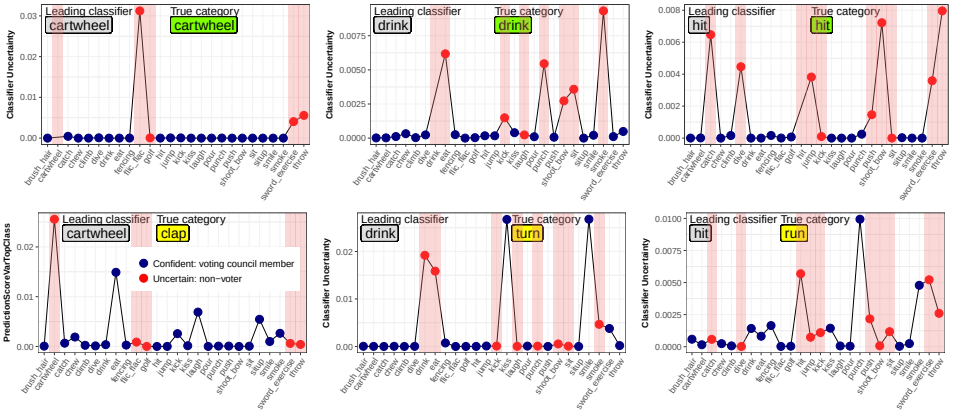

Figure 3: Examples of selective voting for the novelty score of different activities. The first row depicts the case where the samples are of *known* classes and second row for those of *novel* classes. Red points highlight classifiers, which were *excluded* from to the council of the current leader. Their uncertainty is, therefore, ignored when inferring the novelty score.

established datasets, HMDB-51 [15] and UCF-101 [36], for our task[†]. We evenly split each dataset into seen/unseen categories (26/25 for HMDB-51 and 51/50 for UCF-101). Samples of unseen classes will not be available during training, while samples of the remaining set of seen classes is further split into training (70%) and testing (30%) sets, thereby adapting the evaluation framework of [42] for the *generalized* ZS learning scenario. For each dataset, we randomly generate 10 splits and report the average and standard deviation of the recognition accuracy. Using a separate validation split, we optimize the credibility threshold c and compute the threshold for rejection τ for each category as the Equal Error Rate of the ROC.

Architecture details We augment the RGB-stream of the I3D architecture [5] with MC-Dropout. The model is pre-trained on the Kinetics dataset, as described in [5]. The last average pooling is connected to two fully connected layers: a hidden layer of size 256 and the final softmax-classifier layer. These are optimized using SGD with momentum of 0.9, learning rate of 0.005 and dropout probability of 0.7 for 100 epochs. We sample the output scores for $M = 100$ stochastic forward passes applied on the two layers preceding the classifier, while the credibility threshold c is set to 0.001.

Baselines We compare our model to three popular methods for novelty and outlier detection: 1) a One Class SVM [32, 33] with RBF kernel (upper bound on the fraction of training errors ν set to 0.1); 2) a GMM [24, 48] with 8 components; 3) and Softmax probabilities [12, 29] as the value for thresholding. Both SVM and GMM were trained on normalized features obtained from last average pooling layer of I3D pre-trained on the Kinetics dataset [5].

Novelty Detection We evaluate the novelty detection accuracy in terms of a binary classification problem, using the area under curve (AUC) values of the receiver operating characteristic (ROC) and the precision-recall (PR) curves.

We show the robustness of our approach in comparison to the baseline methods in Table 1. All variants of our model clearly outperform the conventional approaches and achieve an ROC-AUC gain of over 7% on both datasets. Along our model variants, *Informed*

[†]Dataset splits used for novelty detection and generalized zero-shot action recognition are provided at https://cvhci.anthropomatik.kit.edu/∼aroitberg/novelty_detection_action_recognition

Novelty Detection Model	HMDB-51		UCF-101	
	ROC AUC %	PR AUC %	ROC AUC %	PR AUC %
Baseline Models				
One-class SVM	54.09 (±3.0)	77.86 (±4.0)	53.55 (±2.0)	78.57 (±2.4)
Gaussian Mixture Model	56.83 (±4.2)	78.40 (±3.6)	59.21 (±4.2)	79.50 (±2.2)
Conventional NN Confidence	67.58 (±3.3)	84.21 (±3.0)	84.28 (±1.9)	93.92 (±0.7)
Our Proposed Model based on Bayesian Uncertainty				
Dictator	71.78 (±1.8)	86.81 (±2.5)	91.43 (±2.3)	96.72 (±1.0)
Uninformed Democracy	73.81 (±1.7)	87.83 (±2.3)	92.13 (±1.8)	97.15 (±0.7)
Informed Democracy	**75.33 (±2.7)**	**88.66 (±2.3)**	**92.94 (±1.7)**	**97.52 (±0.6)**

Table 1: Novelty detection results evaluated as area under the ROC and PR-curves for identifying previously unseen categories (mean and standard deviation over ten dataset splits).

Democracy has proven to be the most effective strategy for novelty score voting, outperforming the *Dictator* by 5.5% and 1.4%, while *Uninformed Democracy* achieved second-best results. We believe that smaller differences in performance gain on the UCF-101 data are due to the much higher supervised classification accuracy on this dataset. Since the categories of UCF-101 are easier to distinguish visually and the confusion is low, there is more agreement between the neurons in terms of their confidence.

Generalized Zero-Shot Learning (GZSL) Next, we evaluate our approach in the context of GZSL, where our novelty detection model serves as a filter to distinguish whether the observed example should be classified with the I3D model in the standard classification setup, or mapped to one of the unknown classes via a ZSL model. We compare two prominent ZSL methods: ConSE [23] and DeViSE [7]. The ConSE model starts by predicting probabilities of the seen classes, and then takes the convex combination of word embeddings of the top K most possible seen classes and select its nearest neighbor from the novel classes in the *word2vec* space. For DeViSE, we train a separate model to regress *word2vec* representations from the visual features. We use the publicly available *word2vec* model that is trained on Google News articles [18].

For consistency, we first report the results for the standard ZS case (*i.e.* U→U) and further extend to the generalized case (*i.e.* U→U+S and U+S→U+S) as shown in Table 2. In the more realistic GZSL setup, our model is not restricted to any group of target labels and is evaluated on actions of seen and unseen category using the *harmonic mean* of accuracies for seen and unseen classes as proposed by [44]. Table 2 shows a clear advantage of employing novelty detection as part of a GZSL framework. While failure of the original ConSE and DeViSE models might be surprising at first glance, such performance drops have been discussed in previous work on ZSL for image recognition [44] and is due to the fact that both models are biased towards labels that were used during training. Our *Informed Democracy* model yields the best recognition rates in every setting and can therefore be indeed successfully applied for multi-label action classification in case of new activities.

5 Conclusion

We introduce a new approach for novelty detection in action recognition. Our model leverages the estimated uncertainty of the category classifiers to detect samples from novel categories not encountered during training. This is achieved by selecting a council of classifiers for each leader (*i.e.* the most confident classifier). The council will validate the decision

Zero-Shot Approach	HMDB-51			UCF-101		
	U→U	U→U+S	U+S→U+S	U→U	U→U+S	U+S→U+S
Standard ConSe Model	21.03 (±2.07)	0 (±0)	0 (±0)	17.85 (±1.95)	0.07 (±0.10)	0.13 (±0.20)
Standard Devise Model	17.27 (±2.01)	0.26 (±0.37)	0.52 (±0.73)	14.48 (±1.13)	0.81 (±0.36)	1.61 (±0.71)
ConSe + Novelty Detection						
One-class SVM	21.03 (±2.07)	10.99 (±1.83)	17.40 (±2.41)	17.85 (±1.95)	10.37 (±1.59)	16.55 (±1.91)
Gaussian Mixture Model	21.03 (±2.07)	13.30 (±2.58)	19.91 (±3.32)	17.85 (±1.95)	9.31 (±1.30)	15.98 (±1.99)
Conventional NN Confidence	21.03 (±2.07)	10.96 (±0.87)	18.56 (±1.22)	17.85 (±1.95)	12.19 (±1.72)	20.91 (±2.59)
Informed Democracy (ours)	21.03 (±2.07)	**13.67 (±1.31)**	**22.27 (±1.79)**	17.85 (±1.95)	**13.62 (±1.94)**	**23.42 (±2.97)**
Devise + Novelty Detection						
One-class SVM	17.27 (±2.01)	8.92 (±1.89)	14.67 (±2.74)	14.48 (±1.13)	8.65 (±1.59)	14.25 (±2.00)
Gaussian Mixture Model	17.27 (±2.01)	10.61 (±2.22)	16.72 (±3.1)	14.48 (±1.13)	7.26 (±0.84)	12.88 (±1.40)
Conventional NN Confidence	17.27 (±2.01)	8.68 (±1)	15.17 (±1.56)	14.48 (±1.13)	10.08 (±1.59)	17.69 (±2.33)
Informed Democracy (ours)	17.27 (±2.01)	**10.73 (±1.47)**	**18.18 (±2.21)**	14.48 (±1.13)	**11.03 (±1.42)**	**19.48 (±2.21)**

Table 2: Accuracy for GZS action recognition with the proposed novelty detection model. U→U: test set consists of unseen actions, the prediction labels are restricted to the unseen labels (standard). U→U+S: test set consists of unseen actions, both unseen and seen labels are possible for prediction. U+S→U+S: generalized ZSL case, both unseen and seen categories are among the test examples and in the set of possible prediction labels (harmonic mean of the seen and unseen accuracies reported.)

made by the leader through voting. Hence, either confirming the classification decision for a sample of a known category or revoking the leader decision and deeming the sample to be novel. We show in a thorough evaluation on two challenging benchmark, that our model outperforms the state-of-the-art in novelty detection. Furthermore, we demonstrate that our model can be easily integrated in a generalized zero-shot learning framework. Combining our model with off-the-shelf zero-shot approaches leads to significant improvements in classification accuracy.

Acknowledgements This work has been partially funded by the German Federal Ministry of Education and Research (BMBF) within the PAKoS project.

References

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 476–483. IEEE, 2017.

[3] MF Augusteijn and BA Folkert. Neural network classification and novelty detection. *International Journal of Remote Sensing*, 23(14):2891–2902, 2002.

[4] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[8] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th International Conference on Learning Representations (ICLR) workshop track*, 2016.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning (ICML-17)*, 2017.

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 18–22, 2018.

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[14] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *arXiv:1705.07115*, May 2017.

[15] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering âĂŸ12*, pages 571–582. Springer, 2013.

[16] Juncheng Liu, Zhouhui Lian, Yi Wang, and Jianguo Xiao. Incremental kernel null space discriminant analysis for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 792–800, 2017.

[17] Markos Markou and Sameer Singh. A neural network-based novelty detector for image sequence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1664–1677, 2006.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] Thomas Moerland, Aswin Chandarr, Maja Rudinac, and Pieter P Jonker. Knowing what you don't know-novelty detection for action recognition in personal robots. In *VISIGRAPP (4: VISAPP)*, pages 317–327, 2016.

[20] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry crowds: detecting violent events in videos. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016.

[21] Reza Mohammadi-Ghazi, Youssef M Marzouk, and Oral Büyüköztürk. Conditional classifiers and boosted conditional gaussian mixture model for novelty detection. *Pattern Recognition*, 2018.

[22] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[23] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[24] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[25] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[26] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Proc. CVPR*, 2017.

[27] Sebastian Ramos, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother. Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pages 1025–1032. IEEE, 2017.

[28] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[29] Charles Richter and Nicholas Roy. Safe visual navigation via deep learning and novelty detection. In *Proc. of the Robotics: Science and Systems Conference*, 2017.

[30] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[31] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11): 2317–2324, 2014.

[32] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

[33] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[35] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[38] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *arXiv preprint arXiv:1801.04264*, 2018.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[40] Gul Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[41] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[42] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124(3):356–383, 2017.

[43] Graham Williams, Rohan Baxter, Hongxing He, Simon Hawkins, and Lifang Gu. A comparative study of rnn for outlier detection in data mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 709–712. IEEE, 2002.

[44] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394*, 2017.

[45] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 63–67. IEEE, 2015.

[46] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, pages 1–25, 2017.

[47] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. 2018.

[48] F Zorriassatine, A Al-Habaibeh, RM Parkin, MR Jackson, and J Coy. Novelty detection for practical pattern recognition in condition monitoring of multivariate processes: a case study. *The International Journal of Advanced Manufacturing Technology*, 25(9-10):954–963, 2005.