SRAC: Self-Reflective Risk-Aware Artificial Cognitive Models for Robot Response to Human Activities

Hao Zhang[1], Christopher Reardon[2], Fei Han[1], and Lynne E. Parker[2]

***Abstract*— In human-robot teaming, interpretation of human actions, recognition of new situations, and appropriate decision making are crucial abilities for cooperative robots ("co-robots") to interact intelligently with humans. Given an observation, it is important that human activities are interpreted the same way by co-robots as human peers so that robot actions can be appropriate to the activity at hand. A novel interpretability indicator is introduced to address this issue. When a robot encounters a new scenario, the pretrained activity recognition model, no matter how accurate in a known situation, may not produce the correct information necessary to act appropriately and safely in new situations. To effectively and safely interact with people, we introduce a new generalizability indicator that allows a co-robot to self-reflect and reason about when an observation falls outside the co-robot's learned model. Based on topic modeling and the two novel indicators, we propose a new *Self-reflective Risk-aware Artificial Cognitive* (SRAC) model, which allows a robot to make better decisions by incorporating robot action risks and identifying new situations. Experiments both using real-world datasets and on physical robots suggest that our SRAC model significantly outperforms the traditional methodology and enables better decision making in response to human behaviors.**

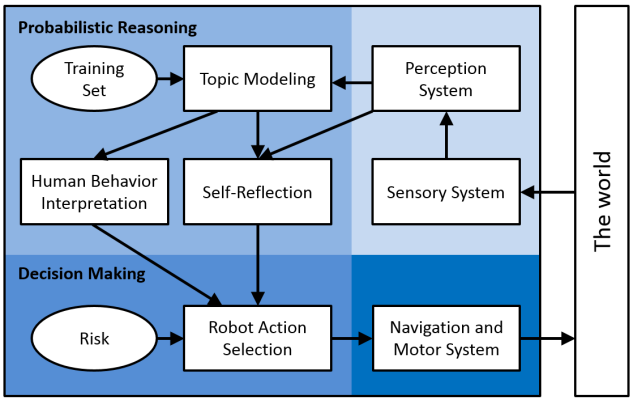


Fig. 1. Overview of the SRAC model for robot response to human activities. The novel self-reflection module allows a co-robot to reason about when the learned knowledge no longer applies. Decisions are made by considering both human activity category distributions and robot action risks. Entities in ellipses are prior knowledge to the SRAC model. Information flows from modules with lighter colors to those with darker colors.

I. Introduction

Recognition of human behaviors and appropriate decision-making are crucial capabilities for a cooperative robot ("co-robot") to understand and interact with human peers. To this end, an intelligent co-robot requires an artificial cognitive model to integrate perception, reasoning, and decision making in order to effectively respond to humans. Artificial cognition has its origin in cybernetics; its intention is to create a science of mind based on logic [1]. Among other mechanisms, cognitivism is a most widely used cognitive paradigm [2]. Several cognitive architectures were developed within this paradigm, including ACT-R [3], Soar [4], C4 [5], and architectures for robotics [6]. Because an architecture represents the connection and interaction of different cognitive components, it cannot accomplish a specific task on its own without specifying each component that can provide knowledge to the cognitive architecture. The combination of the cognitive architecture and components is usually referred to as a cognitive model [2].

Implementing such an artificial cognitive system is challenging, since the high-level processes (e.g., reasoning and decision making) must be able to seamlessly work with the low-level components, e.g., perception, under significant uncertainty in a complex environment [7]. In the context of human-robot collaboration, perceiving human behaviors is a necessary component, where uncertainty arises due to human behavior complexity, including variations in human motions and appearances, and challenges of machine vision, such as lighting changes and occlusion. This perception uncertainty is addressed in this work using the bag-of-visual-words (BoW) representation based on local spatio-temporal features, which has previously shown promising performance [8], [9], [10].

To further process the perceptual data, a high-level reasoning component is necessary for a co-robot to make decisions. In recent years, topic modeling has attracted increasing attention in human behavior discovery and recognition due to its ability to generate a distribution over activities of interest, and its promising performance using BoW representations in robotics applications [10], [11]. However, previous work only aimed at human behavior understanding; the essential task of incorporating topic modeling into cognitive decision making (e.g., selecting a response action) is not well analyzed.

Traditional activity recognition systems typically use accuracy as a performance metric [12]. Because the accuracy metric ignores the distribution of activity categories, which is richer and more informative than a single label, it is not appropriate for decision making. For example, in a task of behavior understanding with two categories, assume that two recognition systems obtain two distributions $[0.8, 0.2]$ and

[1]H. Zhang and F. Han are with the Human-Centered Robotics Lab in the Department of Computer Science and Electrical Engineering, Colorado School of Mines, Golden, CO 80401, USA, hzhang@mines.edu.

[2]C. Reardon and L. E. Parker are with the Distributed Intelligence Lab in the Department of Computer Science and Electrical Engineering, University of Tennessee, Knoxville, TN 37996, USA.

$[0.55, 0.45]$ on a given observation, and the ground truth indicates the first category is correct. Although both systems are accurate, in the sense that the most probable category matches the ground truth, the first model obviously performs better, since it better separates the correct from the incorrect assignment. Previous studies did not consider this important phenomenon.

In real-world applications, artificial cognitive models must be applied in an online fashion. If a co-robot is unable to determine whether its knowledge is accurate, then if it observes a new human behavior that was not presented during the training phase, it cannot be correctly recognized, because the learned behavior recognition model no longer applies. Decision making based on incorrect recognition in situations like these can result in inappropriate or even unsafe robot action response. Thus, an artificial cognitive model requires the capability to self-reflect whether the learned activity recognition system becomes less applicable, analogous to human self-reflection on learned knowledge, when applied in a new unstructured environment. This problem was not well investigated in previous works.

In this paper, we develop a novel artificial cognitive model, based on topic models, for robot decision making in response to human behaviors. Our model is able to incorporate human behavior distributions and take into account robot action risks to make more appropriate decisions (i.e., risk-aware). Also, our model is able to identify new scenarios when the learned recognition subsystem is less applicable (i.e., self-reflective). Accordingly, we call our model the *self-reflective, risk-aware artificial cognitive* (SRAC) model.

Our primary contributions are twofold:

- Two novel indicators are proposed. The *interpretability indicator* (I_I) enables a co-robot to interpret category distributions in a similar manner to humans. The online *generalizability indicator* (I_G) measures the human behavior recognition model's generalization capacity (i.e., how well unseen observations can be represented by the learned model).
- A novel artificial cognitive model (i.e., SRAC) is introduced based on topic models and the indicators, which is able to consider robot action risks and perform self-reflection to improve robot decision making in response to human activities in new situations.

The rest of the paper is organized as follows. We describe the artificial cognitive architecture and its functional modules in Section II. Then, Section III introduces the new indicators. Section IV presents self-reflective risk-aware decision making. Experimental results are discussed in Section V. Finally, we conclude our paper in Section VI.

II. Topic Modeling for Artificial Cognition

A. Cognitive Architecture Overview

The proposed SRAC model is inspired by the C4 cognitive architecture [5]. As shown in Fig. 1, our model is organized into four modules by their functionality:

- *Sensory and perception*: Visual cameras observe the environment. Then, the perception system builds a BoW representation from raw data, which can be processed by topic models.
- *Probabilistic reasoning*: Topic models are applied to reason about human activities, which are trained offline and used online. The training set is provided as a prior that encodes a history of sensory information. This module uses the proposed indicators to select topic models that better match human's perspective, and to discover new activities in an online fashion.
- *Decision making*: Robot action risk based on topic models and the evaluation results is estimated and a response robot action that minimizes this risk is selected. The risk is provided as a prior to the module.
- *Navigation and motor system*: The selected robot action is executed in response to human activities.

B. Topic Modeling

Latent Dirichlet Allocation (LDA) [13], which showed promising activity recognition performance in our prior work [10], is applied in the SRAC model.

Given a set of observations $\{\boldsymbol{w}\}$, LDA models each of K activities as a multinomial distribution of all possible visual words in the dictionary $\boldsymbol{D}$. This distribution is parameterized by $\boldsymbol{\varphi} = \{\varphi_{w_1}, \ldots, \varphi_{w_{|\boldsymbol{D}|}}\}$, where φ_w is the probability that the word w is generated by the activity. LDA also represents each $\boldsymbol{w}$ as a collection of visual words, and assumes that each word $w \in \boldsymbol{w}$ is associated with a latent activity assignment z. By applying the visual words to connect observations and activities, LDA models $\boldsymbol{w}$ as a multinomial distribution over the activities, which is parameterized by $\boldsymbol{\theta} = \{\theta_{z_1}, \ldots, \theta_{z_K}\}$, where θ_z is the probability that $\boldsymbol{w}$ is generated by the activity z. LDA is a Bayesian model, which places Dirichlet priors on the multinomial parameters: $\boldsymbol{\varphi} \sim \mathrm{Dir}(\boldsymbol{\beta})$ and $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\beta} = \{\beta_{w_1}, \ldots, \beta_{w_{|\boldsymbol{D}|}}\}$ and $\boldsymbol{\alpha} = \{\alpha_{z_1}, \ldots, \alpha_{z_K}\}$ are the concentration hyperparameters.

To understand human behaviors, our model applies Gibbs sampling [14] to compute the *per-observation activity distribution* $\boldsymbol{\theta}$. At convergence, the element $\theta_{z_k} \in \boldsymbol{\theta}$, $k = 1, \ldots, K$, is estimated by:

$$\hat{\theta}_{z_k} = \frac{n_{z_k} + \alpha_{z_k}}{\sum_z (n_z + \alpha_z)} \tag{1}$$

where n_z is the count of a visual word being assigned to an activity z_k in the observation.

III. Interpretability and Generalizability

To improve artificial cognitive modeling, we introduce two novel indicators and discuss their relationship in this section, which are the core of the *Self-Reflection* module in Fig. 1.

A. Interpretability Indicator

We observe that accuracy is not an appropriate assessment metric for robot decision making, since it only considers the most probable human activity category and ignores the others. To utilize the category distribution, which contains much richer information, the *interpretability indicator*, denoted by I_I, is introduced. I_I is able to encode how well topic modeling matches human common sense. Like the

accuracy metric, I_I is an extrinsic metric, meaning that it requires a ground truth to compute. Formally, I_I is defined as follows:

Definition 1 (Interpretability indicator): Given the observation $\boldsymbol{w}$ with the ground truth g and the distribution $\boldsymbol{\theta}$ over $K \geq 2$ categories, let $\boldsymbol{\theta}_s = (\theta_1, \ldots, \theta_{k-1}, \theta_k, \theta_{k+1}, \ldots, \theta_K)$ denote the sorted proportion satisfying $\theta_1 \geq \cdots \geq \theta_{k-1} \geq \theta_k \geq \theta_{k+1} \geq \cdots \geq \theta_K \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$, and let $k \in \{1, \cdots, K\}$ represent the index of the assignment in $\boldsymbol{\theta}_s$ that matches g. The interpretability indicator $I_I(\boldsymbol{\theta}, g) = I_I(\boldsymbol{\theta}_s, k)$ is defined as:

$$I_I(\boldsymbol{\theta}_s, k) = \frac{1}{a}\left(\frac{K-k}{K-1} + \mathbb{1}(k=K)\right)\left(\frac{\theta_k}{\theta_1} - \frac{\theta_{k+\mathbb{1}(k\neq K)}}{\theta_k} + b\right) \quad (2)$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $a = 2$, $b = 1$ are normalizing constants.

The indicator I_I is defined over the per-observation category proportion $\boldsymbol{\theta}$, which takes values in the $(K-1)$-simplex [13]. The sorted proportion $\boldsymbol{\theta}_s$ is computed through sorting $\boldsymbol{\theta}$, which is inferred by topic models. In the definition, the ground truth is represented by its location in $\boldsymbol{\theta}_s$, i.e., the k-th most probable assignment in $\boldsymbol{\theta}_s$ matches the ground truth label. The indicator function $\mathbb{1}(\cdot)$ in Eq. (2) is adopted to deal with the special case when $k = K$.

For an observation in an activity recognition task with K categories, given its ground truth index k and sorted category proportion $\boldsymbol{\theta}_s$, we summarize I_I's properties as follows:

Proposition 1 (I_I's properties): The interpretability indicator $I_I(\boldsymbol{\theta}, g) = I_I(\boldsymbol{\theta}_s, k)$ satisfies the following properties:

1. If $k = 1$, $\forall \boldsymbol{\theta}_s$, $I_I(\boldsymbol{\theta}_s, k) \geq 0.5$.
2. If $k = K$, $\forall \boldsymbol{\theta}_s$, $I_I(\boldsymbol{\theta}_s, k) \leq 0.5$.
3. $\forall \boldsymbol{\theta}_s$, $I_I(\boldsymbol{\theta}_s, k) \in [0, 1]$.
4. $\forall k \in \{1, \ldots, K\}$ and $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}'_s$ such that $\theta_1 \geq \theta'_1$, $\theta_k = \theta'_k$ and $\theta_{k+\mathbb{1}(k\neq K)} = \theta'_{k+\mathbb{1}(k\neq K)}$, $I_I(\boldsymbol{\theta}_s, k) \leq I_I(\boldsymbol{\theta}'_s, k)$ holds.
5. $\forall k \in \{1, \ldots, K\}$ and $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}'_s$ such that $\theta_{k+\mathbb{1}(k\neq K)} \geq \theta'_{k+\mathbb{1}(k\neq K)}$, $\theta_1 = \theta'_1$ and $\theta_k = \theta'_k$, $I_I(\boldsymbol{\theta}_s, k) \leq I_I(\boldsymbol{\theta}'_s, k)$ holds.
6. $\forall k \in \{1, \ldots, K\}$ and $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}'_s$ such that $\theta_k \geq \theta'_k$, $\theta_1 = \theta'_1$ and $\theta_{k+\mathbb{1}(k\neq K)} = \theta'_{k+\mathbb{1}(k\neq K)}$, $I_I(\boldsymbol{\theta}_s, k) \geq I_I(\boldsymbol{\theta}'_s, k)$ holds.
7. $\forall k, k' \in \{1, \ldots, K\}$ such that $k \leq k' < K$ and $\forall \boldsymbol{\theta}_s$, $\boldsymbol{\theta}'_s$ such that $\theta_k = \theta'_k$, $\theta_1 = \theta'_1$ and $\theta_{k+\mathbb{1}(k\neq K)} = \theta'_{k+\mathbb{1}(k\neq K)}$, $I_I(\boldsymbol{\theta}_s, k) \geq I_I(\boldsymbol{\theta}'_s, k')$ holds.

Proof: In the supplementary material. ■

The indicator I_I is able to quantitatively measure how well topic modeling can match human common sense, because it captures three essential considerations to simulate the process of how humans evaluate the category proportion $\boldsymbol{\theta}$:

- A topic model performs better, in general, if it obtains a larger θ_k (Property 6). In addition, a larger θ_k generally indicates θ_k is closer to the beginning in $\boldsymbol{\theta}_s$ and further away from the end (Property 7).
  Example: A topic model obtaining the sorted proportion $[0.4, \boxed{0.35}, 0.15, 0.10]$ performs better than a model obtaining $[0.4, \boxed{0.30}, 0.15, 0.15]$, where the ground truth is marked with a box, i.e., $k = 2$ in the example.
- A smaller difference between θ_k and θ_1 indicates better modeling performance (Properties 4 and 5), in general. Since the resulting category proportion is sorted, a small difference between θ_k and θ_1 guarantees θ_k has an even smaller difference from θ_2 to θ_{k-1}.
  Example: A topic model obtaining the sorted proportion $[0.4, \boxed{0.3}, 0.2, 0.1]$ performs better than the model with the proportion $[0.5, \boxed{0.3}, 0.2, 0]$.
- A larger distinction between θ_k and θ_{k+1} generally indicates better modeling performance (Properties 5 and 6), since it better separates the correct assignment from the incorrect assignments with lower probabilities.
  Example: A topic model obtaining the sorted proportion $[0.4, \boxed{0.4}, 0.1, 0.1]$ performs better than the topic model obtaining the proportion $[0.4, \boxed{0.4}, 0.2, 0]$.

Algorithm 1: Left-to-right $Pvwp$ estimation

Input : $\boldsymbol{w}$ (observation), $\mathcal{M}$ (trained topic model), and R (number of particles)
Output : $Pvwp(\boldsymbol{w}|\mathcal{M})$

Initialize $l = 0$ and $N = |\boldsymbol{w}|$;
for *each position* $n = 1$ *to* N *in* $\boldsymbol{w}$ **do**
  Initialize $p_n = 0$;
  for *each particle* $r = 1$ *to* R **do**
    for $n' < n$ **do**
      Sample $z_{n'}^{(r)} \sim P(z_{n'}^{(r)}|w_{n'}, \{\boldsymbol{z}_{<n}^{(r)}\}_{\neg n'}, \mathcal{M})$;
    end
    Compute $p_n = p_n + \sum_t P(w_n, z_n^{(r)} = t|z_{<n}^{(r)}, \mathcal{M})$;
    Sample $z_n^{(r)} \sim P(z_n^{(r)}|w_n, z_{<n}^{(r)}, \mathcal{M})$;
  end
  Update $p_n = \frac{p_n}{R}$ and $l = l + \log p_n$;
end
return $Pvwp(\boldsymbol{w}|\mathcal{M}) \simeq \frac{l}{N}$.

The indicator I_I extends the *accuracy* metric I_A (i.e., rate of correctly recognized data), as described in Proposition 2:

Proposition 2 (Relationship of I_I and I_A): The accuracy measure I_A is a special case of $I_I(\boldsymbol{\theta}_s, k)$, when $\theta_1 = 1.0$, $\theta_2 = \ldots = \theta_K = 0$, and $k = 1$ or $k = K$.

Proof: In the supplementary material. ■

B. Generalizability Indicator

An artificial cognitive model requires the crucial capability of detecting new situations and being aware that the learned knowledge becomes less applicable in an online fashion. To this end, we propose the *generalizability indicator* (I_G), an intrinsic metric that does not require ground truth to compute and consequently can be used online.

The introduction of I_G is inspired by the perplexity metric (also referred to as held-out likelihood), which evaluates a topic model's generalization ability on a fraction of held-out instances using cross-validation [15] or unseen observations [16]. The perplexity is defined as the log-likelihood of words in an observation [17]. Because different observations may contain a different number of visual words, we compute the *Per-Visual-Word Perplexity ($Pvwp$)*. Mathematically, given the trained topic model $\mathcal{M}$ and an observation $\boldsymbol{w}$, $Pvwp$ is defined as follows:

$$Pvwp(\boldsymbol{w}|\mathcal{M}) = \frac{1}{N}\log P(\boldsymbol{w}|\mathcal{M}) = \frac{1}{N}\log \prod_{n=1}^{N} P(w_n|\boldsymbol{w}_{<n}, \mathcal{M}) \quad (3)$$

where $N = |\boldsymbol{w}|$ is the number of visual words in $\boldsymbol{w}$, and the subscript $<n$ denotes positions before n. Because $P(\boldsymbol{w}|\mathcal{M})$ is a probability that satisfies $P(\boldsymbol{w}|\mathcal{M}) \leq 1$, it is guaranteed $Pvwp(\boldsymbol{w}|\mathcal{M}) \leq 0$. The left-to-right algorithm, presented in Algorithm 1, is used to estimate $Pvwp$, which is an accurate and efficient Gibbs sampling method to estimate perplexity [17]. The algorithm decomposes $P(\boldsymbol{w}|\mathcal{M})$ in an incremental, left-to-right fashion, where the subscript $\neg n$ is a quantity that excludes data from the nth position. Given observations $\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$, $Pvwp(\mathcal{W}|\mathcal{M})$ is defined as the average of each observation's perplexity:

$$Pvwp(\mathcal{W}|\mathcal{M}) = \frac{1}{M}\sum_{m=1}^{M} Pvwp(\boldsymbol{w}_m|\mathcal{M}) \tag{4}$$

Based on $Pvwp$, the generalizability indicator I_G, on previously unseen observations in the testing phase or using the held-out instances in cross-validation, is defined as follows:

Definition 2 (Generalizability indicator): Let $\mathcal{M}$ denote a trained topic model, $\mathcal{W}_{\text{valid}}$ denote the validation dataset that is used in the training phase, and $\boldsymbol{w}$ be an previously unseen observation. We define the generalizability indicator:

$$I_G(\boldsymbol{w}) = \begin{cases} \dfrac{\exp(Pvwp(\boldsymbol{w}|\mathcal{M}))}{c \cdot \exp(Pvwp(\mathcal{W}_{valid}|\mathcal{M}))} \\ \quad \text{if } \exp(Pvwp(\boldsymbol{w}|\mathcal{M})) < c \cdot \exp(Pvwp(\mathcal{W}_{valid}|\mathcal{M})) \\ 1 \quad \text{if } \exp(Pvwp(\boldsymbol{w}|\mathcal{M})) \geq c \cdot \exp(Pvwp(\mathcal{W}_{valid}|\mathcal{M})) \end{cases} \tag{5}$$

where $c \in [1, \infty)$ is a constant encoding novelty levels.

We constrain I_G's value in the range $(0, 1]$, with a greater value indicating less novelty, which means an observation can be better encoded by the training set and the topic model generalizes better on this observation. The constant c in Eq. (5) provides the flexibility to encode the degree to which we consider an observation to be novel.

The indicator I_G provides our SRAC model with the ability to evaluate how well a new observation is represented by the training data. Since it is impractical, often impossible, to define an *exhaustive* training set, mainly because some of the categories may not exist at the time of training, the ability to discover novelty and be aware that the learned model is less applicable is essential for safe, adaptive decision making.

C. Indicator Relationship

While the interpretability indicator interprets human activity distributions in a way that is similar to human reasoning, the generalizability indicator endows a co-robot with the self-reflection capability. We summarize their relationship in the cases when a training set is exhaustive (i.e., training contains all possible categories) and non-exhaustive (i.e., new human behavior occurs during testing), as follows:

Observation (Relationship of I_G and I_I): Let $\mathcal{W}_{train}$ be the training dataset used to train a topic model, and I_I and I_G be the model's interpretability and generalizability indicators.

- If $\mathcal{W}_{train}$ is exhaustive, then $I_G \to 1$ and I_I is generally independent of I_G.
- If $\mathcal{W}_{train}$ is non-exhaustive, then I_G takes values that are much smaller than 1; I_I also takes small values and is moderately to strongly correlated with I_G.

This observation answers the critical question of whether a better generalized topic model can lead to better recognition performance. Intuitively, if $\mathcal{W}_{train}$ is non-exhaustive and a previously unseen observation $\boldsymbol{w}$ belongs to a novel category, which is indicated by a small I_G value, a topic model trained on $\mathcal{W}_{train}$ cannot accurately classify $\boldsymbol{w}$. On the other hand, if $\boldsymbol{w}$ belongs to a category that is known in $\mathcal{W}_{train}$, then $I_G \to 1$ and the recognition performance over $\boldsymbol{w}$ only depends on the model's performance on the validation set used in the training phase. The meaning and relationship of the indicators I_I and I_G are summarized in Table I, where the gray area denotes that it is generally impossible for a topic model to obtain a low generalizability but a high interpretability, as a model is never correct when presented with a novel activity.

TABLE I

MEANING AND RELATIONSHIP OF I_I AND I_G. THE GRAY AREA DENOTES THAT THE SITUATION IS GENERALLY IMPOSSIBLE.

	I_G: low	I_G: high
I_I: low	Category is novel Model is *not* applicable	Category is *not* novel Model is *not* well interpreted
I_I: high		Category is *not* novel Model is well interpreted

IV. SELF-REFLECTIVE RISK-AWARE DECISION MAKING

Another contribution of this research is a decision making framework that is capable of incorporating activity category distribution, robot self-reflection (enabled by the indicators), and co-robot action risk, which is realized in the module of *Decision Making* in Fig. 1. Our new self-reflective risk-aware decision making algorithm is presented in Algorithm 2.

Given the robot action set $\boldsymbol{a} = \{a_1, \ldots, a_S\}$ and the human activity set $\boldsymbol{z} = \{z_1, \ldots, z_K\}$, an action-activity risk r_{ij} is defined as the amount of discomfort, interference, or harm that can be expected to occur during the time period if the robot takes a specific action $a_i, \forall i \in \{1, \ldots, S\}$ in response to an observed human activity $z_j, \forall j \in \{1, \ldots, K\}$. While $\boldsymbol{\theta}$ and I_G are computed online, the risks $\boldsymbol{r} = \{r_{ij}\}_{S \times K}$, with each element $r_{ij} \in [0, 100]$, are manually estimated off-line by domain experts and used as a prior in the decision making module. In practice, the amount of risk is categorized into a small number of risk levels for simplicity's sake. To assign a value to r_{ij}, a risk level is first selected. Then, a risk value is determined within that risk level. As listed in Table II, we define four risk levels with different risk value ranges in our application. We intentionally leave a five-point gap between critical risk and high risk to increase the separation of critical risk from high risk actions.

TABLE II

RISK LEVELS AS PRIOR KNOWLEDGE TO OUR COGNITIVE MODEL.

Levels	Values	Definition
Low risk	[1,30]	Unsatisfied with the robot's performance.
Medium risk	[31,60]	Annoyed or upset by the robot's actions.
High risk	[61,90]	Interfered with, interrupted, or obstructed.
Critical risk	[95,100]	Injured or worse (i.e., a safety risk).

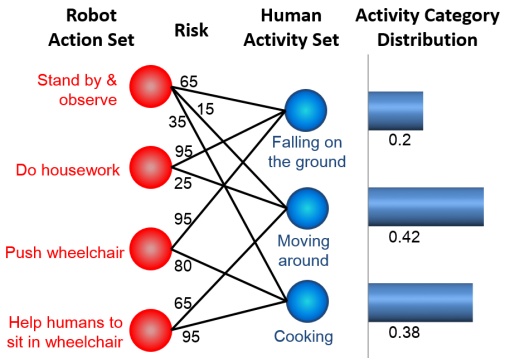


Fig. 2. An illustrative example of a bipartite network (left) and the per-observation activity distribution (right) in assistive robotics applications.

A bipartite network $\mathcal{N}=\{\boldsymbol{a},\boldsymbol{z},\boldsymbol{r}\}$ is proposed to graphically illustrate the risk matrix $\boldsymbol{r}$ of robot actions $\boldsymbol{a}$ associated with human activities $\boldsymbol{z}$. In this network, vertices are divided into two disjoint sets $\boldsymbol{a}$ and $\boldsymbol{z}$, such that every edge with a weight r_{ij} connects a vertex $a_i \in \boldsymbol{a}$ to a vertex $z_j \in \boldsymbol{z}$. An example of such a bipartite network is depicted in Fig. 2 for assistive robotics applications. The bipartite network also has a tabular representation (for example, in Table III). Given the bipartite network, for a new observation $\boldsymbol{w}$, after $\boldsymbol{\theta}$ and $I_G(\boldsymbol{w})$ are computed in the probabilistic reasoning module, the robot action $a^\star \in \boldsymbol{a}$ is selected according to:

$$a^\star = \underset{a_i:i=1,\ldots,S}{\arg\min} \left(\frac{1-I_G(\boldsymbol{w})}{K} \cdot \sum_{j=1}^{K} r_{ij} + I_G(\boldsymbol{w}) \cdot \sum_{j=1}^{K} (\theta_j \cdot r_{ij}) \right) \quad (6)$$

The risk of taking a specific robot action is determined by two separate components: activity-independent and activity-dependent action risks. The activity-independent risk (that is $\frac{1}{K}\sum_{j=1}^{K} r_{ij}$) measures the inherent risk of an action, which is independent of the human activity context information, i.e., computing this risk does not require the category distribution. For example, the robot action "standing-by" generally has a smaller risk than "moving backward" in most situations. The activity-dependent risk (that is $\sum_{j=1}^{K}(\theta_j \cdot r_{ij})$) is the average risk weighted by context-specific information (i.e., the human activity distribution). The combination of these two risks is controlled by I_G, which automatically encodes preference over robot actions. When the learned model generalizes well over $\boldsymbol{w}$, i.e., $I_G(\boldsymbol{w}) \to 1$, the decision making process prefers co-robot actions that are more appropriate to the recognized human activity. Otherwise, if the model generalizes poorly, indicating new human activities occur and the learned model is less applicable, our decision making module would ignore the recognition results and select co-robot actions with lower activity-independent risk.

V. Experiments

To evaluate the performance of the proposed SRAC model, we use three real-world benchmark human activity datasets: the Weizmann [18], KTH [19], and UTK3D datasets [10]. We also demonstrate our cognitive model's effectiveness in a human following task using a real autonomous mobile robot.

Algorithm 2: Self-reflective risk-aware decision making

Input : $\boldsymbol{w}$ (observation), $\mathcal{M}$ (trained topic model), and $\mathcal{N}$ (decision making bipartite network)
Output : $a^\star$ (Selected robot action with minimum risk)

Estimate per-observation activity proportion $\boldsymbol{\theta}$ of $\boldsymbol{w}$;
Compute generalizability indicator $I_G(\boldsymbol{w})$;
for *each robot action* $i = 1$ *to* S **do**
  Estimate activity-independent risk: $r_i^{in} = \frac{1}{K}\sum_{j=1}^{K} r_{ij}$;
  Calculate activity-dependent risk: $r_i^{de} = \sum_{j=1}^{K}(\theta_j \cdot r_{ij})$;
  Combine activity-independent and dependent risks, and assign to per-observation action risk vector: $\boldsymbol{r}^a(i) = (1 - I_G(\boldsymbol{w})) \cdot r_i^{in} + I_G(\boldsymbol{w}) \cdot r_i^{de}$;
end
Select optimal robot action $a^\star$ with minimum risk in $\boldsymbol{r}^a$;
return $a^\star$.

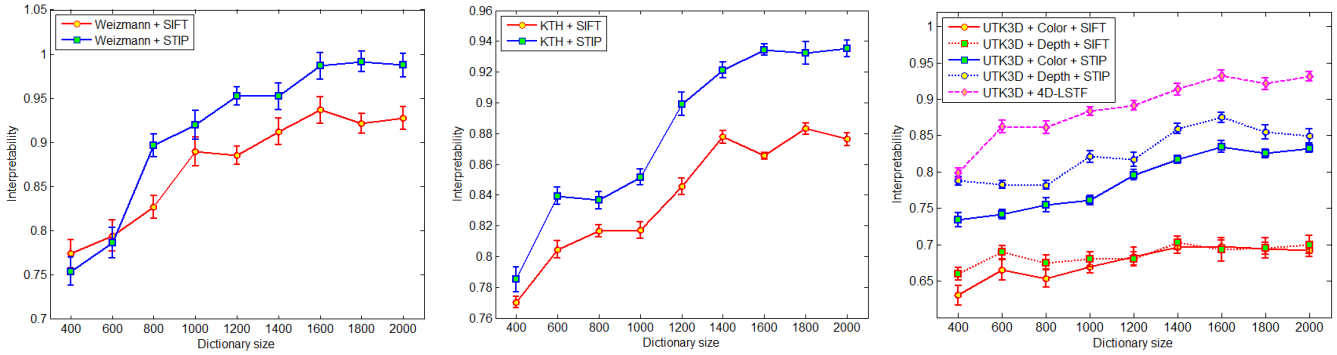
We validate our SRAC model over multiple standard visual features, including the space-time interest points (STIP) [20], histogram of oriented gradients (HOG) [19], and histogram of optical flow (HOF) features [19] for color or depth data, as well as the 4-dimensional local spatio-temporal features (4D-LSTF) [10] for RGB-D data. The k-means approach is applied to construct a dictionary and convert the features to a BoW representation for each observation [10].

A. Activity Recognition

We first evaluate the SRAC model's capability to recognize human activities using the interpretability indicator I_I, when the training set is exhaustive. In this experiment, each dataset is split into disjoint training and testing sets. We randomly select 75% of data instances in each category as the training set, and employ the rest of the instances for testing. During training, fourfold cross-validation is used to estimate model parameters. Then, the interpretability of the topic model is computed using the testing set, which is fully represented by the training set and does not contain novel human activities. This training-testing process is repeated five times to obtain reliable results.

Experimental results of the interpretability and its standard deviation versus the dictionary size are illustrated in Fig. 3. Our SRAC model obtains promising recognition performance in terms of interpretability: 0.989 is obtained using the STIP feature and a dictionary size 1800 on the Weizmann dataset, 0.952 using the STIP feature and a dictionary size 2000 on the KTH dataset, and 0.936 using the 4D-LSTF feature and a dictionary size 1600 on the UTK3D dataset. In general, STIP features perform better than SIFT features for color data, and 4D-LSTF features perform the best for RGB-D visual data. The dictionary size in the range [1500, 2000] can generally result in satisfactory human activity recognition performance. The results are also very consistent, as illustrated by the small error bars in Fig. 3, which demonstrates our interpretability indicator's consistency.

The model's interpretability is also evaluated over different activity categories using the UTK3D dataset, which includes more complex activities (i.e., sequential activities) and contains more information (i.e., depth). It is observed that topic



(a) Weizmann dataset (b) KTH dataset (c) UTK3D dataset

Fig. 3. Variations of model interpretability and its standard deviation versus dictionary size using different visual features over benchmark datasets.

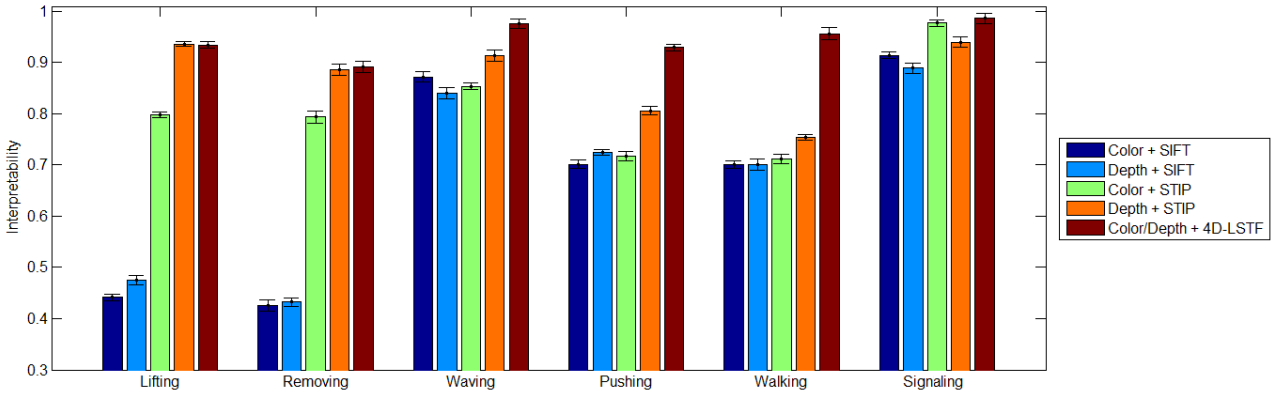


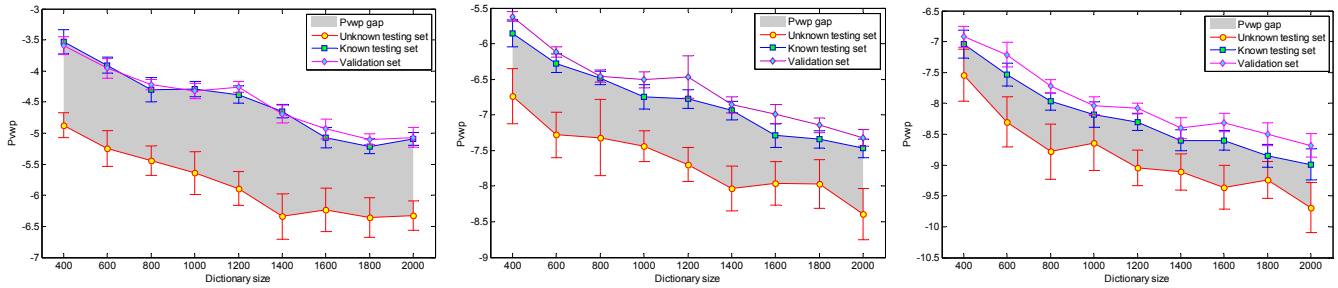
Fig. 4. Model interpretability over the activities in the UTK3D dataset using different features and a dictionary size of 1600.

modeling's interpretability varies for different activities. This performance variation is affected by three main factors: the topic model's modeling ability, feature and BoW's representability, and human activity complexity and similarity. For example, since the LDA topic model and SIFT features are not capable of modeling time, the reversal human activities including "lifting a box" and "removing a box" in the UTK3D dataset cannot be well distinguished, as illustrated in Fig. 4. Since sequential activities (e.g., "removing a box") are more complex than repetitive activities (e.g., "waving"), they generally result in low interpretability. Since "pushing" and "walking" are similar, which share motions such as moving forward, they can also reduce interpretability. This observation provides general guidance for designing future recognition systems with the SRAC model.

B. Knowledge Discovery

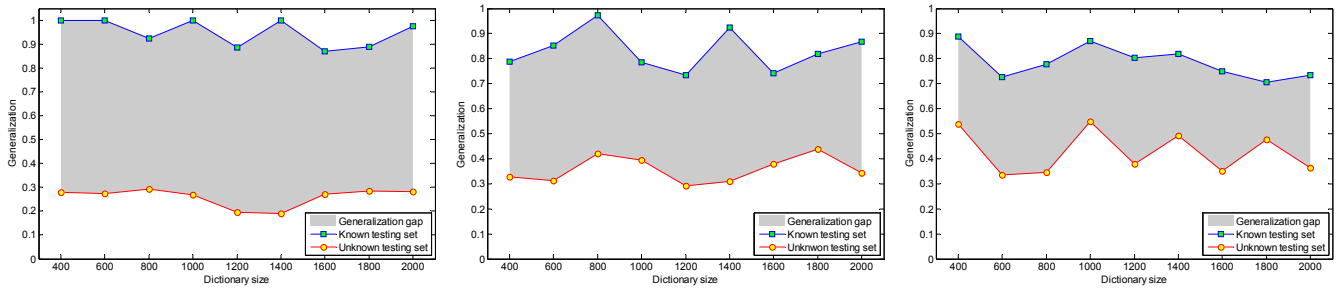
We evaluate the SRAC model's capability to discover new situations using the generalizability indicator I_G, when the training dataset is non-exhaustive (i.e., new human activities occur during testing). A non-exhausted setup is created by dividing the used benchmark datasets as follows. We place all data instances of one activity in the *unknown testing set*, and randomly select 25% of the instances from the remaining activities in the *known testing set*. The remaining instances are placed in the training set for learning, based on fourfold cross-validation. To evaluate the model's ability to discover each individual human activity, given a dataset that contains K activity categories, the experiments are repeated K times, each using one category as the novel activity. Visual features that achieve the best model interpretability over each dataset are used in this set of experiments i.e., STIP features for the Weizmann and KTH datasets and 4D-LSTF features for the UTK3D dataset.

Variations of $Pvwp$ values versus the dictionary size over the validation set (in cross-validation), known testing set, and unknown testing set are shown in Fig. 5. Several important phenomena are observed. First, there exists a large $Pvwp$ gap between the known and unknown testing sets, as shown by the gray area in the figure, indicating that topic models generalize differently over data instances from known and unknown activities. A better generalization result indicates a less novel instance, which can be better represented by the training set. Since data instances from the known testing and validation sets are well represented by the training set, the $Pvwp$ gap between them is small. As shown in Fig. 5(a), it is possible that the known testing set's $Pvwp$ value is greater than the $Pvwp$ value of the validation set, if its data instances can be better represented by the training set. Second, Fig. 5 shows that the gap's width varies over different datasets: the Weizmann dataset generally has the largest $Pvwp$ gap, followed by the KTH dataset, and then the UTK3D dataset.



(a) Weizmann dataset + STIP features (b) KTH dataset + STIP features (c) UTK3D dataset + 4D-LSTF features

Fig. 5. Variations of topic modeling's $Pvwp$ versus dictionary size over validation set, known and unknown testing sets.



(a) Weizmann dataset (b) KTH dataset (c) UTK3D dataset

Fig. 6. Variations of our model's generalizability versus dictionary size over known and unknown testing sets for all datasets.

The gap's width mainly depends on the observation's novelty, in terms of the novel activity's similarity to the activities in the training dataset. This similarity is encoded by the portion of overlapping features. A more novel activity is generally represented by a set of more distinct visual features with less overlapping with the features existing during training, which generally results in a larger gap. For example, activities in the Weizamann dataset share fewer motions and thus contain a less number of overlapping features, which leads to a larger gap. Third, when the dictionary size increases, the model's $Pvwp$ values decrease at a similar rate. This is because in this case, the probability of a specific codeword appearing in an instance decreases, resulting in a decreasing $Pvwp$ value.

The generalizability indicator I_G's characteristics are also empirically validated on the known and unknown testing sets, as illustrated in Fig. 6. An important characteristic of I_G is its invariance to dictionary size. Because $Pvwp$ over testing and validation sets has similar decreasing rate, the division operation in Eq. (5) removes the variance to dictionary size. In addition, a more novel activity generally leads to a smaller I_G value. For example, the Weizmann dataset has a smaller I_G value over the unknown testing set, because its activities are more novel in the sense that they share less overlapping motions. In general, we observe I_G is smaller than 0.5 for unknown activities and greater than 0.7 for activities that are included in training sets. As indicated by the gray area in Fig. 6, similar to $Pvwp$, there exists a large gap between the I_G values over the unknown and known testing datasets. The average I_G gap across different dictionary sizes is 0.69 for the Weizmann dataset, 0.48 for the KTH dataset, and 0.36 for the UTK3D dataset. This reasoning process, based on I_G,

(a) Falling (b) Turning (c) Walking

Fig. 7. Experiment setup for validating the SRAC model's decision making ability in a human following task using a Turtlebot 2 robot.

provides a co-robot with the critical self-reflection capability, and allows a robot to reason about when new situations occur as well as when the learned model becomes less applicable.

C. Decision Making

We assess our SRAC model's decision making capability using a Turtlebot 2 robot in a human following task, which is important in many human-robot teaming applications. In this task, a robotic follower needs to decide at what distance to follow the human teammate. We are interested in three human behaviors: "walking" in a straight line, "turning," and "falling." With perfect perception and reasoning, i.e., a robot always perfectly interprets human activities, we assume the ideal robot actions are to "stay far from the human" when he or she is walking in a straight line (to not interrupt the human), "move close to the human" when the subject is turning (to avoid losing the target), and "stop beside the human" when he or she is falling (to provide assistance).

In order to qualitatively assess the performance, we collect 20 color-depth instances from each human behaviors to train

TABLE III

THE RISK MATRIX USED IN THE ROBOT FOLLOWING TASK.

Robot Actions	Falling	Turning	Walking
Stay besides humans	0	20	50
Move close	90	0	20
Stay far away	95	80	0

the SRAC model, using a BoW representation based on 4D-LSTF features. The risk matrix used in this task is presented in Table III. We evaluate our model in two circumstances. `Case 1`: exhaustive training (i.e., no unseen human behaviors occur in testing). In this case, the subjects only perform the three activities during testing with small variations in motion speed and style. `Case 2`: non-exhaustive training (i.e., novel movements occur during testing). In this case, the subjects not only perform the activities with large variations, but also add additional movements (such as jumping and squatting) which are not observed in the training phase. During testing, each activity is performed 40 times. The model performance is measured using failure rate, i.e., the percentage with which the robot fails to stop besides to help the human or loses the target.

Experimental results are presented in Table IV, where the traditional methodology, which selects the co-robot actions only based on the most probable human activity, is used as a baseline for comparison. We observe that the proposed SRAC model significantly decreases the failure rate in both exhaustive and non-exhaustive setups. When the training set is exhaustive and no new activities occur during testing (`Case 1`), the results demonstrate that incorporating human activity distributions and robot action risks improves decision making performance. When the training set is non-exhaustive and new activities occur during testing (`Case 2`), the SRAC model significantly outperforms the baseline model. In this situation, if I_G has a very small value, according to Eq. 6, our model tends to select safer robot actions, i.e., "stay beside humans," since its average risk is the lowest, which is similar to the human common practice "playing it safe in uncertain times." The results show the importance of self-reflection for decision making especially under uncertainty.

TABLE IV

FAILURE RATE (%) IN EXHAUSTIVE (`CASE 1`) AND NON-EXHAUSTIVE (`CASE 2`) EXPERIMENTAL SETTINGS.

Exp. settings	Models	Fail to assist	Fail to follow
Exhaustive (`Case 1`)	Baseline	10.5%	15%
	SRAC	0.5%	5.5%
Non-exhaustive (`Case 2`)	Baseline	45.5%	60%
	SRAC	24.5%	35.5%

VI. CONCLUSION

In this paper, we propose a novel self-reflective risk-aware artificial cognitive model based on topic modeling. Two new indicators are introduced and combined in the SRAC model. The interpretability indicator generalizes the accuracy metric and enables a robot to interpret category distributions in a similar fashion to humans. By considering this distribution, our model is also able to incorporate robot action risks. The generalizability indicator measures how well an observation can be represented by the learned knowledge, which allows for self-reflection that can enable the SRAC model to identify new scenarios. Through incorporating robot action risks (by reasoning about category distributions) and the self-reflection ability (realized by the generalizability indicator), our model makes better decisions that can more appropriately respond to human behaviors, which is validated in experiments both using benchmark datasets and on real robots.

REFERENCES

[1] F. J. Varela and J. Dupuy, *Understanding Origins*, ch. Whence perceptual meaning? A cartography of current ideas, pp. 235–263. Kluwer Academic Publishers, 1992.

[2] D. Vernon, G. Metta, and G. Sandini, "A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents," *IEEE Transactions on Evolutionary Computation*, vol. 11, pp. 151–180, Apr. 2007.

[3] J. R. Anderson, "ACT: A simple theory of complex cognition," *American Psychologist*, vol. 51, pp. 355–365, Apr. 1996.

[4] J. E. Laird, A. Newell, and P. S. Rosenbloom, "SOAR: an architecture for general intelligence," *Artificial Intelligence*, vol. 33, pp. 1–64, Sept. 1987.

[5] D. Isla, R. Burke, M. Downie, and B. Blumberg, "A layered brain architecture for synthetic creatures," in *International Joint Conferences on Artificial Intelligence*, 2001.

[6] C. Burghart, R. Mikut, R. Stiefelhagen, T. Asfour, H. Holzapfel, P. Steinhaus, and R. Dillmann, "A cognitive architecture for a humanoid robot: a first approach," in *IEEE-RAS International Conference on Humanoid Robots*, 2005.

[7] U. Schmid, M. Ragni, C. Gonzalez, and J. Funke, "The challenge of complexity for cognitive systems," *Cognitive Systems Research*, vol. 12, no. 3-4, pp. 211–218, 2011.

[8] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[9] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[10] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition.," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[11] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 2013.

[12] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *National Academy of Sciences*, 2004.

[15] C. C. Musat, J. Velcin, S. Trausan-Matu, and M.-A. Rizoiu, "Improving topic evaluation using conceptual knowledge," in *International Joint Conference on Artificial Intelligence*, 2011.

[16] D. Blei and J. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, 2006.

[17] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *International Conference on Machine Learning*, 2009.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247–2253, Dec. 2007.

[19] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.