Symbolic Knowledge Representation for Video Analysis: A Representational Systems Theory Approach

Paper #5887

Abstract. This paper presents RST-Vision, a novel approach to video behaviour analysis that integrates Representational Systems Theory (RST) with computer vision models, focusing on violence detection in surveillance scenarios. Existing approaches typically rely on deep learning models, which do not generalise well across datasets, and lack interpretability and explicit reasoning capabilities. Our method transforms visual scenes into high-level structural graphs through RST-based symbolic representation, encoding spatial and temporal relationships between detected entities and structures. The framework processes video key frames using pose estimation models to extract anatomical features, which are then evaluated through predefined RST-based expert schemas for behaviour classification and detection. Experimental results across various datasets exhibit superior discrimination capability in terms of accuracy and ROC AUC figures, with the same symbolic model and no retraining required. Notably, our approach demonstrates enhanced performance with increasing pose estimation model quality. This suggests that RST-Vision models fundamental behavioural patterns effectively, rather than relying on dataset-specific features. Our findings indicate that symbolic architectures like RST-Vision hold significant promise for advancing more transparent and robust video and image analysis.

1 Introduction

The detection of violent behaviour in video footage plays a vital role in maintaining public safety and preventing criminal activities. Although modern computer vision approaches leveraging neural networks have demonstrated impressive accuracy rates [15], they present two significant limitations. Firstly, these systems operate as black boxes, lacking both interpretability and explicit reasoning mechanisms. Secondly, they exhibit poor generalisation across different datasets, as their dependence on dataset-specific visual features often results in overfitting. This becomes particularly problematic when the models encounter out-of-distribution scenarios, where their performance deteriorates substantially.

We present RST-Vision, a purely symbolic framework that employs Representational Systems Theory (RST) [5] to achieve complete interpretability by transforming visual scenes into explicit high-level structural representations. The framework implements a unified entity representation approach, converting skeletal data from pose estimation models into a topology tree-like graph structures. This representation encompasses anatomical features, including limb positions and joint angles, which are encoded and evaluated through the RST framework's structural transformation mechanism [11]. Beyond skeletal data, the framework accommodates additional features such as dimensional measurements, chromatic features, and environmental characteristics and objects. Through its user-friendly interface, RST-Vision enables domain experts to define explicit rules (expert schemas) for violent event detection, automatically generating alerts when these conditions are met. The framework's sophisticated logical reasoning capabilities support the implementation of complex rule sets, allowing for nuanced event detection and classification.

Through skeletal representations that abstract human figures from their surroundings, our approach captures movement dynamics whilst mitigating the risk of overfitting to background elements in training data, thereby enhancing generalisation capabilities. Finally, our evaluation demonstrates how rigorous logical rule reasoning achieves high accuracy and precision. We particularly examine the framework's discriminative capabilities, minimising false positive rates, which is a critical metric for real-world deployment.

RST-Vision presents several contributions:

- We propose a novel symbolic knowledge representation approach based on representational systems theory, for interpretable video behaviour analysis. Competitive performance metrics in accuracy, precision and ROC AUC are observed, compared to state-of-the-art neural models, while preserving complete interpretability through the use of symbolic rules.
- RST-Vision achieves enhanced cross-dataset generalisation, indicating successful modelling of general skeletal patterns rather than dataset-specific features.
- We demonstrate that the primary factor for enhancing performance is the quality of the underlying pose estimation. The clear linear relationship between accurate object representation and violence detection accuracy suggests that as pose estimation models improve, there is significant potential for advancements of violence detection precision.

2 Background and Related Work

Our methodology is grounded in two primary components: pose estimation models that facilitate perceptual recognition of raw pixel inputs from real-world video scenes; and a symbolic Representational Systems Theory (RST) that leverages the pose models to provide abstraction and generalisation capabilities for violence scene description and detection.

2.1 Violence Detection Methods

The increasing deployment of surveillance cameras in public spaces presents a growing need for automated systems to monitor them effectively for public safety purposes.

The Flow Gated Network [3] uses the optical flow channel, in addition to RGB channel, as a gating mechanism to determine which temporal information should be preserved or discarded, rather than relying on traditional pooling strategies.

The SepConvLSTM [4] relies on two key information streams: human skeletal data and temporal changes between frames. It employs OpenPose [2] to extract human pose information. For temporal information, the system analyses frame-to-frame differences rather than using more computationally expensive optical flow calculations. It processes video sequences using a Convolutional Long Short-Term Memory (ConvLSTM) network to aggregate temporal information, followed by depthwise separable convolutions for efficient post-processing. In contrast, whilst our RST-Vision does utilise human pose information, our key innovation lies in a novel explicit symbolic representation of human poses. This deterministic approach addresses shortcomings inherent in existing methods that rely on implicit or probabilistic representations.

VadCLIP [17] leverages the pre-trained CLIP vision-language model, and uses only video-level labels available during training. It utilises dual-branch architecture that combines coarse-grained and fine-grained detection approaches. The first branch performs binary classification to determine if a video contains anomalous events, whilst the second branch utilises CLIP's vision-language alignment capabilities to identify specific types of anomalies.

All theses three approaches in violence detection predominantly rely on neural network architectures, which necessitate substantial training datasets. While these models demonstrate some generalisability across different datasets, their performance metrics—specifically accuracy and ROC-AUC scores—decline markedly when trained on one dataset and evaluated on another, indicating significant overfitting to their training data. Our approach marks a departure from data-driven methods, offering a more interpretable and potentially more generalisable solution to violence detection.

2.2 Perceptual Pose Estimation Models

Recent developments in computer vision have yielded several notable human pose estimation models, including OpenPose [2], MediaPipe Pose Landmarker [7], YOLO Pose [8], MMPose [12], RTMPose [6], and SwinPose [18]. These models process raw image or video input to generate coordinate data for human skeletal key points. The generalisability of the RST framework enabled us to successfully integrate all pose models. For our experimental evaluation, we employed RTMPose due to its superior performance in multi-person scenarios—especially in fighting sequences with significant occlusion—while delivering state-of-the-art average precision [6]. RTMPose's provision of confidence scores for each keypoint is required for our quality threshold evaluations, and it conveniently adapts the YOLO Pose format.

2.3 Representational Systems Theory Framework

Representational Systems Theory (RST) is designed to abstractly encode a wide variety of representations under a single, unifying paradigm using *structure graphs*. These capture syntax, entailment and other properties of each representation [11]. RST has proven effective across multiple domains, including mathematical formula processing, diagrammatic reasoning [16], and automated problem-solving [9].

Figure 1 demonstrates two types of representations: one is a numerical expression $1 + 2$, and the other is a dot diagram. They are both constructed in RST formalism and encoded as a structure graph. Figure 1 shows how one dot diagram stacks with another two dots diagram to form a new two layers three dots pile ⁘ , using an RST constructor *stack*.

RST uses *transfer schemas* to define the invariant relation between two types of representations. As shown in Figure 2, it maps the dot diagram in Figure 1 into a numerical expression $1 + 2$. The corresponding constructors are the *infixOp* operator for the numerical expressions and the *stack* constructor for the dot diagrams. The blue arrow indicates preconditions: ∘ is represented as 1, ∞ is represented as 2, and the ∘ and ∞ are disjoint. Then the red arrow indicates the consequent: ⁘ is represented as $1 + 2$.

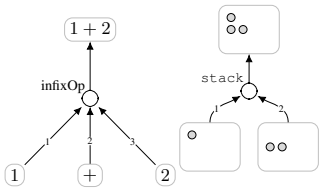

Figure 1: RST structure graphs.

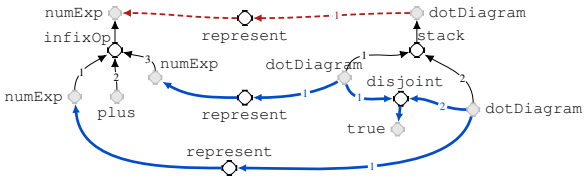

Figure 2: RST transfer schema of expressions and dot diagrams.

RST-Vision extends RST into real-world scene analysis of human poses and environmental context. We use RST structure graphs to encode and generalise all pose estimation models mentioned above, and represent pose data as formal mathematical expressions. We use RST transfer schemas to define general expert rules that systematically transform structure graphs of pose data into perceptual structures and transformations for anomalous event detection. These schemas generalise across different datasets. Built on RST's theoretical foundation, our approach ensures explicit, rigorous, and explainable construction-to-reasoning processes while systematically encoding domain expertise through the transfer schema within a formal framework.

Oruga operationalises the language of RST. It implements data structures of type systems, constructor specifications, constructions and transfer schemas. It consists of an engine for executing transformations using *structure transfer* [10]. We use RST and the *Oruga* language in RST-Vision for the following key advantages:

- General Structure and Visual Representation: unlike many Knowledge Representation and Reasoning(KRR) formalisms, RST is explicitly designed for structured diagrammatic reasoning while maintaining strong expressive generality.
- Reusability and Adaptability: RST enables domain experts to intuitively define and modify rules, facilitating rapid adaptation to new domains.
- Cross-Representation Transformation: RST supports seamless conversion between different representations (e.g., pose models to semantic structures), a critical feature for our framework.

3 Methodology

3.1 Framework Outline

As shown in Figure 3, we propose a four-stage hierarchical processing pipeline that bridges perceptual (vision) and symbolic (reasoning) paradigms in visual understanding systems. The architecture pipeline comprises two primary domains: perceptual intelligence

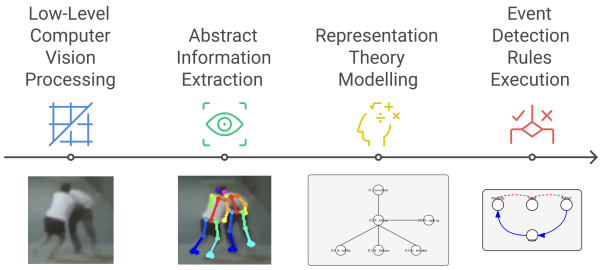

Figure 3: Four-stage visual processing pipeline bridging perceptual and symbolic domains.

```
typeSystem poseDiagram =
    types _:person, _:head, _:upperBody, _:lowerBody,
        _:face, _:leftFace, _:rightFace, _:nose, _:neck, ...
        _:leftBody, _:rightBody,
    order face < head, head < person, upperBody < person, ...

conSpec poseDiagramG:poseDiagram =
    constructors
        joinPose : [head, upperBody, lowerBody] -> person,
        joinHead : [nose, rightFace, leftFace, neck] -> head,
        joinRightFace   : [rightEye, rightEar] -> rightFace,
        joinUpperBody   : [leftBody, rightBody] -> upperBody,
```

Figure 4: Pose type and constructor in the Oruga language (partial).

(stages 1-2) and symbolic intelligence (stages 3-4). The first stage implements computer vision processing through OpenPose for human pose estimation. This foundation enables robust feature extraction from raw visual inputs. Stage two performs abstract information extraction, focusing on geometric primitives (object squares) and anthropometric data (human body key points). This abstraction layer transforms low-level visual features into structured intermediate representations encoding key points in a human pose. The third stage leverages Representation Theory (RST) modelling to bridge the semantic gap between perceptual and symbolic domains. Through RST constructions (structure graphs), RST-Vision converts geometric and spatial relationships into formal symbolic representations. The final stage executes event detection rules in the form of transfer schemas, implementing a verification system that raises event alerts. This phase enables high-level reasoning about complex visual scenarios (e.g., lifting leg for kicking). The systematic progression from raw visual input to symbolic reasoning enables robust event detection while maintaining interpretability as the whole pipeline including types and schemas are all white-box (thanks to symbolic transformation rules) throughout the processing pipeline.

3.2 *Human Pose and Environment Construction*

Our framework employs a variety of pose estimation models, each producing human poses in one of these three distinct formats: OpenPose (25 keypoints), YOLO Pose (16 keypoints), and PoseLandmarker. To manage this heterogeneity, we designed a general framework based on RST constructions, which efficiently handles these diverse representation formats as follows.

Type Systems Within RST, concrete objects are represented as *tokens* with associated *types*. Figure 4 presents a fragment of the type system utilised for a Pose Diagram.

Constructor Specifications We developed an automated system to translate JSON output into RST construction in the Oruga language [10], enabling abstract encoding across multiple representations. The system employs specifically defined constructors, such as *joinPose*, *joinLowerBody*, and *joinLeftLeg*, to construct human poses from keypoint coordinates. As an illustration, the *joinHead* construc-

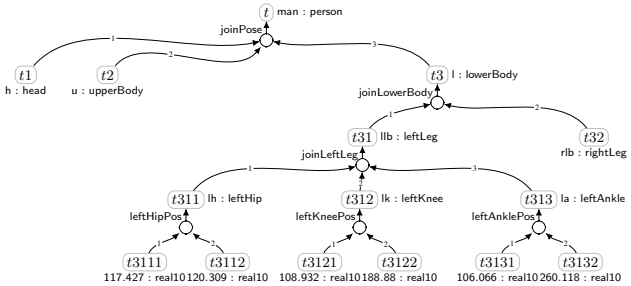

Figure 5: A structure graph of one human pose in RST.

tor functions as a quadruple operator, establishing that a head is composed of four distinct types: nose, rightFace, leftFace, and neck. Figure 4 illustrates the construction specification, which maintains compatibility across all pose models and constructions. Figure 5 demonstrates a specific implementation, presenting a partial example of keypoint annotations for the left leg and complete body structure, encompassing the head, upper body, and lower body components. The comprehensive construction script is available in Listing 1 in the Appendix.

Environmental constructions encompass both human postural data and physical object specifications, including dimensional attributes (e.g., turnstile bar heights). For a detailed implementation of this encoding methodology, please refer to Listing 12 in the Appendix.

Construction Space consists of the type system, the constructor specifications (Figure 4) and the structure graph (Figure 5).

3.3 *Transformation of Pose Structure into Rules*

Transfer schema functions as a set of inference rules for deriving relations across construction spaces. In the Oruga language, these schemas are defined through the specification of source and target patterns, alongside antecedent and consequent constraints. We establish a target space that mirrors the pose construction space, incorporating alert conditions within the antecedent or consequent constraints. Thus, a successful structure transfer indicates that the pose construction has fulfilled the predefined alert conditions.

Within the RST transfer schema framework we encode expert-defined behavioural rules for violence detection. Transfer schemas convert qualitative human observations, such as "arm is bent for punching" or "leg is lifted for kicking", into precise mathematical formulae. These formulae evaluate specific geometric relationships through RST transfer schemas—for instance, "The angle between the neck, shoulder, and arm exceeds $160°$, indicating a punching motion" or "A kicking motion is detected when the angle between the leg and trunk is smaller than $120°$". The implementation encompasses three primary transfer schemas (see Figure 6 for an example):

- One-person punching motion.
- One-person kicking motion.
- Two-person body lean to each other.

3.4 *Open Goals Resolution and Alerts Generation*

The transfer schema creates a mapping between concrete source pose data (containing specific coordinate values) and an abstract pose representation. This schema defines what constitutes abnormal conditions for different pose components by examining the relationships between keypoints and their positions in space. During the structure transfer process, we refer to any unresolved pose constructions as

```
tSchema joinLeftBody:(poseDiagramG, indicator, ...) =
source
    p:?pid:person <- joinPose[
        h:head <- joinHead[
            ...
            nk:neck <- neckPos[xn:real10,yn:real10]
        ],
        ...
    ]
target t:leftPunch, op:?pid:person
antecedent
    :metaTrue <- larger[sx:real10<- threePointsAngle[
            lw:point <- coordinates[xw:real10,yw:real10],
            ls:point <- coordinates[xs:real10,ys:real10],
            ln:point <- coordinates[xn:real10,yn:real10]],
            v:160.0:real10
    ]
consequent a:alert
```

Figure 6: Transfer schema of the punching in the Oruga language.

open goals. The completion of the transfer process occurs when all open goals are resolved. When this happens, it indicates that the specified detection rule criteria have been satisfied, triggering an alert. Figure 7 presents an example of an open goal during the resolution process, illustrating the intermediate state before the angle numerical calculations taking place. The transfer schema utilises specialised functions, namely *ThreePointAngle* and *FourPointAngle*, to compute angles between coordinate point sets.

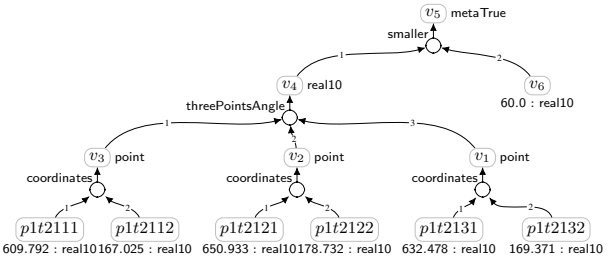

Figure 7: RST goal structure graph showing three-point angle computation from pose data. The goal remains open as the condition "$threePointsAngle < 60^{\circ}$" is not satisfied in the current frame.

3.5 Alert Confidence and Classification Threshold

Let us consider the Real-World Fighting (RWF) dataset as an exemplar case. Each video sequence v comprises t seconds of footage captured at a frame rate of 30 frames per second (fps), yielding a total of $30t$ frames.

For each frame F_i in the sequence, where $i \in \{1, ..., 30t\}$, our pipeline generates a binary classification: $F_i = \{1$ if an alert is raised, and 0 otherwise $\}$.

We define the video alert confidence C as the fourth root of the proportion of frames that trigger an alert: $C = \sqrt[4]{\frac{\sum_{i=1}^{30t} F_i}{30t}}$, where the numerator is the sum of all the alert frames in that video clip, and the denominator is the total number of frames. The fourth root normalisation function transforms confidence scores to a standardised [0,1] interval whilst preserving proportional spacing. Figure 14 in the Appendix illustrates the distributional variations across each root transformation. Given a confidence threshold ϕ, we classify the video sequence as containing fighting activity when $C > \phi$. The threshold ϕ presents a classical precision-sensitivity trade-off: a lower threshold increases the model's sensitivity but reduces precision, whilst a higher threshold enhances precision at the cost of sensitivity. This normalised confidence metric serves as the foundation for computing the Receiver Operating Characteristic (ROC) Area Under Curve (AUC) metrics presented in our experimental evaluations.

4 Evaluation

4.1 Dataset Characteristics

Table 1 summarises the five benchmark datasets used in our evaluation. The datasets represent diverse scenarios of violent behaviour, sourced from different platforms and contexts: Real World Fight (RWF-2000) [3] comprises 2,000 videos from YouTube, evenly split between fighting and non-fighting scenes. RWF-2000 dataset [3] is specifically designed for violence detection in surveillance scenarios. The dataset overcomes limitations of previous collections by offering high-quality footage from authentic surveillance cameras rather than acted scenes or movie clips. Open World Fight (OWF) [14] contains 600 videos collected from TikTok, capturing more casual and spontaneous interactions. Real Life Violence Situations (RLVS) [13] provides 2,000 surveillance-style videos from YouTube, while Hockey Fight (HF) [1] focuses on sport-specific violence with 1,000 clips from hockey games. The datasets feature varying video resolutions, ranging from low-resolution hockey footage (360×288) to high-definition YouTube content (1280×720), enabling evaluation of our method's robustness across different video qualities. This diversity in sources, contexts, and specifications allows for comprehensive assessment of model generalisation capabilities.

4.2 Evaluation Metrics

To evaluate the efficacy of our proposed method, we employ two performance metrics: precision and accuracy. Precision quantifies the model's ability to avoid false positives by measuring the ratio of correctly identified fighting incidents to the total number of fighting alerts generated. Accuracy provides a holistic assessment of the model's performance by measuring the ratio of correct predictions (both fighting and non-fighting) to the total number of videos analysed. We define the key variables used by our evaluation metrics:

- N_f: Number of fighting video clips in the dataset.
- N_n: Number of non-fighting video clips in the dataset.
- A_f: Number of fighting videos triggering alerts as expected.
- A_n: Number of non-fighting videos triggering alerts incorrectly.
- O_f: Number of low quality fighting video clips excluded.
- O_n: Number of low quality non-fighting video clips excluded.

The performance metrics are calculated as:

- True Positive (TP): A_f
- False Negative (FN): $N_f - O_f - A_f$
- False Positive (FP): A_n
- True Negative (TN): $N_n - O_n - A_n$

We define our primary evaluation metrics:

Precision (measures alert accuracy): $P = \frac{A_f}{A_f + A_n}$

Overall Accuracy: $Acc = \frac{A_f + N_n - O_n - A_n}{N_n - O_n + N_f - O_f}$

4.3 Pose Estimation Quality Filtering

The reliability of pose estimation models can be significantly impacted by various environmental factors, including occlusion from other individuals or objects, and subject distance or angle from the camera. These challenges can result in either failed detection or imprecise key point estimation. Given that pose estimation accuracy directly influences our method's effectiveness, implementing robust quality filtering mechanisms is crucial.

To evaluate model performance under varying conditions of pose estimation quality, we established a configurable quality score

Dataset	Abbv.	# Fighting	# Non-fighting	Source	Common Resolution
Real World Fight	RWF-2000	1000	1000	YouTube	1280×720, 320×240
Open World Fight	OWF	300	300	Tik Tok	1280×720, 626×360
Real Life Violence Situations	RLVS	1000	1000	YouTube	224×224, 1280×720
Hockey Fight	HF	500	500	Hockey Games	360×288

Table 1: Datasets details.

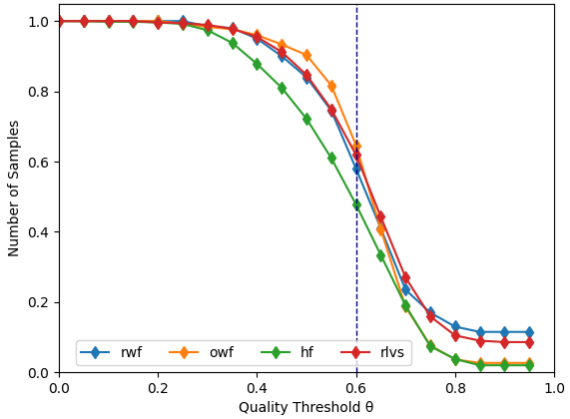

Figure 8: Normalised data retention rates versus quality threshold θ across RWF, OWF, HF, and RLVS datasets. The analysis shows consistent retention above 90% for $\theta \leq 0.4$ and above 50% for $\theta \leq 0.6$ across all datasets. The vertical dashed line indicates the selected operational threshold of $\theta = 0.6$.

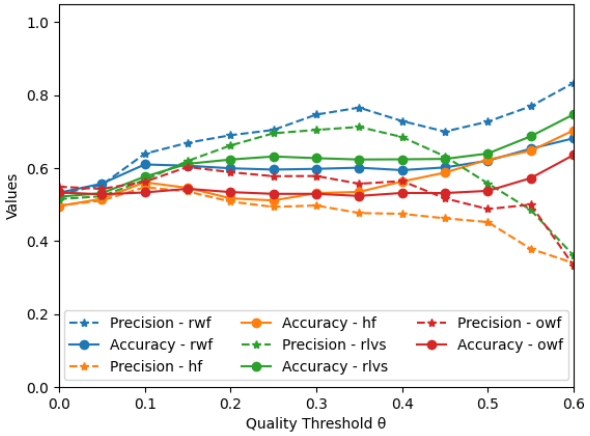

Figure 9: Performance metrics of RST-Vision versus quality score threshold across RWF, OWF, HF, and RLVS datasets. Precision (dashed lines) and accuracy (solid lines) are shown for each dataset, demonstrating gradually increasing accuracy trends. For the RWF dataset, precision consistently increases, while for the other three datasets' precision increases up to $\theta = 0.35$ before declining.

threshold mechanism. This enables systematic analysis of model performance as pose estimation quality improves, providing insights into the relationship between input quality and detection accuracy.

5 Experimental Results

In our experiments, we answer the following questions:

Q1. How do the RST-Vision's precision, accuracy and ROC-AUC performance compare with existing violence detection models?

Q2. What is the relationship between pose estimation quality thresholds and the method's performance?

Q3. What quality filtering threshold achieves the optimal balance between prediction accuracy and data retention for existing datasets?

Through this analysis, we aim to establish robust quality filtering criteria that maximise the method's effectiveness while maintaining practical applicability across diverse scenarios.

5.1 Pose Estimation Quality and Quantity

Figure 8 quantifies the relationship between quality thresholds θ and data retention rates across RWF-500, OWF, HF, and RLVS datasets. The analysis reveals that retention rates are maintained above 90% up to a threshold of $\theta = 0.4$, beyond which they exhibit a significant decrease. As the threshold approaches 0.8, the number of retained video samples decreases significantly, highlighting the inherent trade-off between pose estimation quality and dataset preservation. Based on empirical analysis, we establish a quality threshold of $\theta = 0.60$, which maintains a 50-70% retention rate across datasets while ensuring sufficient pose estimation quality.

5.2 Precision and Accuracy Analysis

The relationship between quality threshold and model performance is illustrated in Figure 9.

Accuracy Comparison with Baselines Our RST-Vision is compared with two existing models SepConvLSTM and Flow Gated Network across four benchmark datasets in Table 2. SepConvLSTM and Flow Gated Network are trained on the RWF training set. RST-Vision is evaluated using a quality score threshold $\theta = 0.60$. SepConvLSTM achieves superior performance on the in-domain RWF test set (0.93) but shows performance degradation on out-of-domain datasets. In contrast, RST-Vision and Flow Gated Network both exhibit consistent performance across domains. These results suggest that skeleton-based models, as discussed in Section 2.1 may offer better generalisation capabilities compared to pure neural-based models for cross-dataset detection.

Quality Threshold Analysis We observe improvement in accuracy metrics as the quality threshold increases up to 0.6, consistently across all datasets. The increasing trend stems from our model's growing confidence as the pose estimation model quality improves, resulting in fewer false positives and higher overall accuracy. The precision for the RWF dataset keeps rising, whereas the precision for the other three datasets increases with the threshold until 0.35.

5.3 ROC-AUC Analysis

To assess the discriminative ability of RST-Vision in binary classification across all decision thresholds, we measure its capacity to distinguish positive from negative samples. Table 3 presents ROC-AUC performance comparison among RST-Vision, SepConvLSTM and Flow Gated Network across four benchmark datasets. While both

Models/Dataset	RWF	HF	RLVS	OWF
SepConvLSTM	**0.93**	0.67	**0.79**	**0.73**
Flow Gated Network	0.67	**0.73**	0.75	0.64
RST-Vision (Ours)	0.68	0.70	0.75	0.64

Table 2: Cross-dataset performance comparison with baselines: model accuracy on in-domain (RWF-2000, shaded) and out-of-domain (HF, RLVS, OWF) datasets. Note that RST-Vision employs predefined expert rules without dataset-specific training.

Models/Dataset	RWF	HF	RLVS	OWF
SepConvLSTM	**0.94**	0.93	**0.93**	0.89
Flow Gated Network	0.79	0.72	0.80	0.79
RST-Vision (Ours)	0.93	**0.95**	0.89	**0.96**

Table 3: ROC-AUC performance comparison with baselines across datasets. Shaded cells indicate in-domain performance on RWF training data. Bold values highlight best performance for each dataset. RST-Vision achieves competitive or superior performance without dataset-specific training.

SepConvLSTM and Flow Gated Network require in-domain training on RWF training data, RST-Vision operates zero-shot without dataset-specific training. Gray-shaded cells denote in-domain evaluation on RWF training data, with bold entries indicating superior performance. RST-Vision achieves top performance on two of the four datasets, and on the RWF dataset, its performance (0.93) closely approaches that of the best in-domain model SepConvLSTM (0.94).

Figure 10 presents the ROC curves comparing RST-Vision and SepConvLSTM models on OWF datasets. A notable characteristic of RST-Vision is its ability to achieve peak True Positive Rate (TPR) at substantially lower False Positive Rate (FPR) values (0.09), as evidenced by the steeper initial curves across all datasets. This pattern is particularly pronounced in the RWF-500 and OWF datasets, where RST-Vision reaches near-maximum TPR (>0.95) at FPR values below 0.2. The AUC scores for both models are competitive, ranging from 0.89 to 0.96, demonstrating strong overall performance. Figure 13 in the Appendix shows ROC Curves for all the datasets. The rapid ascent of RST-Vision's ROC curves indicates superior early detection capabilities, suggesting that the model can identify true positives while maintaining low false alarm rates. This characteristic is particularly valuable in real-world violence detection scenarios where false positives can be costly. In contrast, SepConvLSTM exhibits more gradual curves, requiring higher FPR thresholds to achieve comparable TPR values. The consistency of this pattern across diverse datasets suggests that RST-Vision's efficient discrimination capability is a robust feature of the model architecture rather than a dataset-specific phenomenon.

Comparative ROC-AUC Analysis Across Qualities We analyse the correlation between pose estimation quality thresholds and model discrimination capability through ROC-AUC metrics. Figure 11 illustrates the performance dynamics of RST-Vision and SepConvLSTM architectures evaluated on four benchmark datasets. RST-Vision demonstrates monotonic improvement in discrimination capability as quality thresholds increase, reaching superior AUC values (>0.95) in the high-threshold regime (0.8-1.0). Conversely, SepConvLSTM exhibits threshold-invariant behaviour, maintaining consistent AUC scores across the quality spectrum. These patterns suggest that RST-Vision's performance is particularly sensitive to pose estimation fidelity, with higher quality poses enabling more robust action recognition. When the first two stages achieve high accuracy (near their maximum potential), the rules consistently deliver near-perfect results.

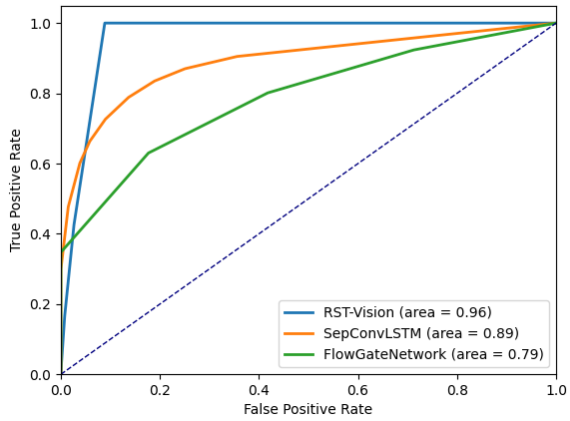

Figure 10: ROC curves for RST-Vision, SepConvLSTM, and Flow Gated Network evaluated on OWF dataset. SepConvLSTM achieves AUC score 0.89, and RST-Vision top AUC at 0.96, and demonstrating steeper initial ascent and earlier TPR saturation. Flow Gated Network shows comparatively lower performance AUC 0.79. The diagonal dotted line indicates random classifier performance.

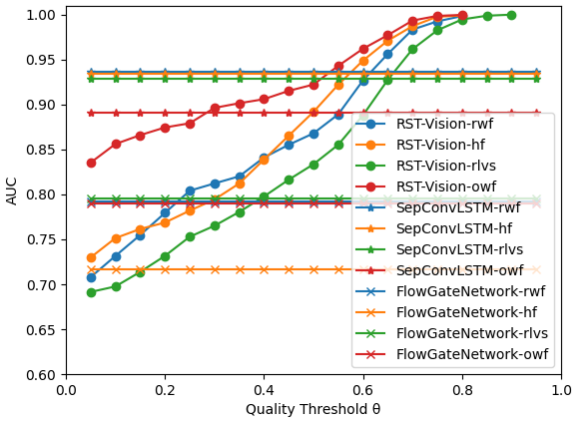

Figure 11: AUC performance of RST-Vision and SepConvLSTM models across varying score thresholds on RWF, OWF, HF, and RLVS datasets. While SepConvLSTM exhibits relatively stable AUC values across thresholds, RST-Vision shows consistent improvement with increasing threshold values, surpassing SepConvLSTM's performance when the score threshold exceeds 0.60. Higher AUC values indicate better model discrimination capability at each threshold.

Answers to Research Questions

Q1. The results in Section 5.2 show RST-Vision offers comparable accuracy with the other models. In Section 5.3 RST-Vision achieves equal or superior ROC-AUC on three out of four datasets.

Q2. Section 5.3 demonstrates RST-Vision ROC-AUC increases further as quality thresholds increase.

Q3. Section 5.2 and Section 5.1 suggest an optimal threshold of 0.6 for balancing quality and quantity.

5.4 Generalising to Fare Evasion Detection

```
graph x:poseDiagramG =
e:env <- joinEnv[
    t:man:person <- joinPose[
        t1:h:head <- joinHead[
        ......
    ],
    b0:turnstile:bar <- barHeight[b1:157.5:real10]
  ]
```

Figure 12: Environmental representation in RST-Vision: pose construction with turnstile bar height parameter in Oruga code.

Cases	Bent Leg	Under Bar	Over Bar	Normal	Total
Correct/Total	5/5	3/3	2/2	2/2	12/12

Table 4: Performance analysis of RST-Vision on fare evasion detection: perfect classification accuracy across all test categories.

We extended RST-Vision to address fare evasion detection, which encompasses the identification of individuals circumventing ticket barriers without presenting valid tickets—whether by vaulting over or manoeuvring beneath the turnstiles—using our carefully curated dataset comprising 12 exemplars that include both transgressive behaviours and legitimate passage through the barriers, sourced from publicly available video footage. This scenario shift demonstrates a key advantage of our approach: while neural-based models would require extensive dataset reconstruction and retraining, RST-Vision required only minor modifications to expert rules while maintaining its core framework. This experiment also illustrates how RST-Vision can incorporate environmental objects (such as turnstile bar height) alongside human pose data.

The fundamental components of RST-Vision—types, pose constructions, and transfer schema templates—remained unchanged from the violence detection task. To enable under/over bar detection, we simply modelled the turnstile environment with a new constructor called "barHeight" representing the vertical position of the barrier, as shown in Figure 12. This environmental parameter can be automatically determined using Hough Line Transformation and simple expert rules.

Our implementation utilised three primary transfer schemas, (illustrated in Listing 5 in the Appendix):

- Bent leg posture: Hip-Knee-Ankle three point angle measuring less than 120 degrees;
- Under-barrier evasion: Shoulder position detected below the turnstile bar along the vertical axis;
- Over-barrier jumping: Detection of any foot component (Ankle, Heel, Toe) positioned higher than the bar's vertical position;
- Normal passage: Subject exhibiting none of the above characteristic patterns.

Table 4 presents the results of our fare avoidance detection evaluation. RST-Vision achieved perfect accuracy across all test cases, correctly identifying all instances of bent leg postures (5/5), under-bar evasions (3/3), over-bar jumps (2/2), and normal passage (2/2), for a total performance of 12/12 correct classifications. These results demonstrate RST-Vision's exceptional adaptability to new scenarios with minimal modification requirements.

6 Discussion and Limitations

Uncertainty Our framework addresses uncertainty through two key mechanisms:

- Pose Modelling Stage: The initial two stages explicitly incorporate uncertainty tolerance. The pose models are designed to accommodate variations in real-world scenes, generating an abstracted representation suitable for subsequent rule-based processing.
- Rule-Based Processing: The rule model inherently handles uncertainty by focusing on strict criteria identified through key frames in the video timeline. This approach naturally filters ambiguous or marginal scenes, providing robustness against uncertainty.

Notably, while our current rule-based implementation demonstrates optimal performance, the framework does not preclude future integration of probabilistic rule models that may offer enhanced uncertainty handling for complex real-world scenarios.

Enhancement The case analysis shows false positives alerts triggered by non-violent movements resembling violence and false negatives arising from partially obscured violence. RST-Vision's strong performance at higher pose estimation quality thresholds underscores the critical importance of accurate object representation in achieving high accuracy for this domain. Conversely, higher thresholds reduce sample size, potentially impacting neural network models' training quality. In contrast to other models such as SepConvLSTM which utilise pose model information but do not demonstrate increasing performance alongside rising quality, this finding underscores the unique strengths of RST-Vision's design.

Generalisability The RST-Vision framework offers exciting possibilities for future development. Its generalisability suggests its potential applicability to other anomaly detection scenarios such as burglary, riot situations, shoplifting, and vandalism. This would involve modifying existing RST Oraga schemas or creating new ones tailored to each specific application. The core architecture remains the same thanks to the generality of RST. The RST framework ensures full reusability of types, constructors, schemas, and reasoning tactics. Domain experts can easily adapt existing templates to new scenarios—for example, repurposing fight detection schemas for fare evasion by implementing rules that evaluate joint angles between shoulder, hip, and knee to identify turnstile-jumping postures.

Connections While many neural-symbolic models enforce strong coupling between their neural and symbolic components, we argue that such tight integration is not strictly necessary for high performance. A key innovation of our approach is the decoupling of these two components, allowing the symbolic rules to function as a crisp, well-defined core model. This architectural choice also enables the experiments in Figure 11, where we identify the sub-symbolic component as the primary performance bottleneck—a novel and significant finding in this domain.

7 Conclusions and Further Work

This research highlights RST-Vision's symbolic architecture's efficacy in tackling the vital issue of automated violence detection from video. Its inherent expressiveness and demonstrably superior generalisation across datasets demonstrate its potential as a robust surveillance solution.

Improved pose estimation techniques could enable near-perfect accuracy for RST-Vision. Its logical operation-based nature allows potential computational efficiency gains for reduced hardware complexity, making it suitable for low-end devices and real-time applications. This positions RST-Vision as a valuable approach for real-world surveillance to enhance safety.

Our current model focuses on individual video frame snapshots, but could be extended to incorporate temporal analysis across multiple sequential frames to better capture gesture trajectories and movement patterns, potentially enhancing detection accuracy.

Ethical Statement

The public datasets used in this research aim to enhance public safety through a robust, interpretable violence detection method, with ethical considerations accounted for.

References

[1] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar. Violence Detection in Video Using Computer Vision Techniques. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, editors, *Computer Analysis of Images and Patterns*, volume 6855, pages 332–339. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23677-8 978-3-642-23678-5. doi: 10.1007/978-3-642-23678-5_39. URL http://link.springer.com/10.1007/978-3-642-23678-5_39. Series Title: Lecture Notes in Computer Science.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, May 2019. URL http://arxiv.org/abs/1812.08008. Issue: arXiv:1812.08008 arXiv:1812.08008 [cs].

[3] M. Cheng, K. Cai, and M. Li. RWF-2000: An Open Large Scale Video Database for Violence Detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190, Jan. 2021. doi: 10.1109/ICPR48806.2021.9412502. URL https://ieeexplore.ieee.org/abstract/document/9412502. ISSN: 1051-4651.

[4] G. Garcia-Cobo and J. C. SanMiguel. Human skeletons and change detection for efficient violence detection in surveillance videos. *Computer Vision and Image Understanding*, 233:103739, Aug. 2023. ISSN 1077-3142. doi: 10.1016/j.cviu.2023.103739. URL https://www.sciencedirect.com/science/article/pii/S1077314223001194.

[5] M. Jamnik and P. Cheng. Endowing machines with the expert human ability to select representations: Why and how. *Human-Like Machine Intelligence*, pages 355–378, 2021. doi: 10.1093/oso/9780198862536.003.0018. ISBN: 9780198862536.

[6] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose, July 2023. URL http://arxiv.org/abs/2303.07399. Issue: arXiv:2303.07399 arXiv:2303.07399 [cs].

[7] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. MediaPipe: A Framework for Building Perception Pipelines, June 2019. URL http://arxiv.org/abs/1906.08172. Issue: arXiv:1906.08172 arXiv:1906.08172 [cs].

[8] D. Maji, S. Nagori, M. Mathew, and D. Poddar. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In *YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss*, pages 2637–2646, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/ECV/html/Maji_YOLO-Pose_Enhancing_YOLO_for_Multi_Person_Pose_Estimation_Using_Object_CVPRW_2022_paper.html.

[9] D. Raggi, G. Stapleton, A. Stockdill, M. Jamnik, G. G. Garcia, and P. C. Cheng. How to (Re)represent it? In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, volume 2020-Novem, pages 1224–1232. IEEE Computer Society, Nov. 2020. ISBN 978-1-7281-9228-4. doi: 10.1109/ICTAI50040.2020.00185. ISSN: 10823409.

[10] D. Raggi, G. Stapleton, M. Jamnik, A. Stockdill, G. G. Garcia, and P. C. Cheng. Oruga: An Avatar of Representational Systems Theory. *CEUR Workshop Proceedings*, 3227:1–5, 2022. ISSN 16130073. ISBN: 0000000265676.

[11] D. Raggi, G. Stapleton, M. Jamnik, A. Stockdill, G. G. Garcia, and P. C.-H. Cheng. Representational Systems Theory: A Unified Approach to Encoding, Analysing and Transforming Representations. *CoRR*, 2022. URL http://arxiv.org/abs/2206.03172. arXiv: 2206.03172.

[12] A. Sengupta, F. Jin, R. Zhang, and S. Cao. mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs. *IEEE Sensors Journal*, 20(17):10032–10044, Sept. 2020. ISSN 1558-1748. doi: 10.1109/JSEN.2020.2991741. URL https://ieeexplore.ieee.org/abstract/document/9083948?casa_token=qvf9NC2E-0sAAAAA:viX-ZO5lfNTR3Voz32nLoxRYY7A244H5R6YSbprrRTWuEPG4_64Z3-gPKZqbjo1-feBwD5cwjDo. Number: 17 Conference Name: IEEE Sensors Journal.

[13] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab. Violence Recognition from Videos using Deep Learning Techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85, Dec. 2019. doi: 10.1109/ICICIS46948.2019.9014714. URL https://ieeexplore.ieee.org/document/9014714/?arnumber=9014714.

[14] M. Tan, Z. Cheng, Y. Li, D. Nanqing, Y. Xiaoyun, and T. Lukasiewicz. Open World Fights: A strong benchmark for evaluating generalization capabilities, Aug. 2024. URL https://github.com/ResearchPaperCodes/AMP-OWF. Issue: owf original-date: 2024-08-08T12:43:52Z.

[15] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Comput. Surv.*, 55(10):200:1–200:44, Feb. 2023. ISSN 0360-0300. doi: 10.1145/3561971. URL https://dl.acm.org/doi/10.1145/3561971.

[16] L. Wu, S. Choi, D. Raggi, A. Stockdill, G. G. Garcia, F. Colarusso, P. C. Cheng, and M. Jamnik. Generation of Visual Representations for Multi-Modal Mathematical Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23850–23852, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/30586. Issue: 21.

[17] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang. VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):6074–6082, Mar. 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i6.28423. URL https://ojs.aaai.org/index.php/AAAI/article/view/28423. Number: 6.

[18] Z. Xiong, C. Wang, Y. Li, Y. Luo, and Y. Cao. Swin-Pose: Swin Transformer Based Human Pose Estimation, June 2022. URL http://arxiv.org/abs/2201.07384. Issue: arXiv:2201.07384 arXiv:2201.07384 [cs].