Meeting in the Middle: A Top-down and Bottom-up Approach to Detect Pedestrians

Affan Shaukat, Andrew Gilbert, David Windridge and Richard Bowden
CVSSP, University of Surrey, Guildford, UK
{*A.Shaukat, A.Gilbert, D.Windridge, R.Bowden*}*@surrey.ac.uk*

Abstract

This paper proposes a generic approach combining a bottom-up (low-level) visual detector with a top-down (high-level) fuzzy first-order logic (FOL) reasoning framework in order to detect pedestrians from a moving vehicle. Detections from the low-level visual corner based detector are fed into the logical reasoning framework as logical facts. A set of FOL clauses utilising fuzzy predicates with piecewise linear continuous membership functions associates a fuzzy confidence (a degree-of-truth) to each detector input. Detections associated with lower confidence functions are deemed as false positives and blanked out, thus adding top-down constraints based on global logical consistency of detections. We employ a state of the art visual detector on a challenging pedestrian detection dataset, and demonstrate an increase in detection performance when used in a framework that combines bottom-up detections with (fuzzy FOL-based) top-down constraints.

1. Introduction

A huge amount of work is being conducted in the area of pedestrian detection, especially from moving vehicles. The ability to detect people in images is required for a number of important applications ranging from surveillance and robotics, to intelligent automotive vehicles. The goal is rendered difficult due to large variations in human pose and clothing, as well as varying backgrounds and environmental conditions.

Most of the current approaches to this task of pedestrian detection have treated this as a recognition task [11] and hence used generic object detection and recognition techniques to solve the problem. Also there have been attempts to use other detector types such as infra-red or LIDAR based point clouds. However, these detector based hardware approaches can struggle especially with medium and far scale pedestrians.

We propose an approach to fuse a bottom-up state of the art visual detection system and a top-down logic reasoning framework. This allows far greater performance than either alone. In the next section, we introduce the visual detection system, with Section 3 and 4 describing the top down reasoning framework. Results on the fusion of the approaches are shown in Section 5 on a challenging dataset, before conclusions are drawn.

2. Visual Pedestrian Detection

There are many approaches of pedestrian detection through the use of image descriptors, including HOG [2], or Gabor filters [3]. These approaches all use single frames, with no temporal information. However, as has been found within action recognition [3, 5], the use of spatio-temporal features allows for greater performance on dynamic actions such as walking pedestrians. Therefore, a spatio-temporal based corner feature descriptor is used to provide additional temporal information. The approach is based on the approach by Gilbert and Bowden [5], we adapt their approach to detect the *walking* action of the pedestrians.

2D corners are detected in the three orthogonal planes of the video sequence (x, y), (x, t) and (y, t). There are a large number of corners detected per frame giving an over-complete set of features with large amounts of redundancy and noise. Each corner is encoded as a three-digit number denoting the spatio-temporal plane in which it was detected, the scale at which it was detected, and its orientation. These corners are then used within an iterative hierarchical grouping process to form descriptive compound features. Each corner is grouped within a cuboid-based neighbourhood. A set of grouped corners is called a *Transaction* and these are collected to form a *Transaction database*. This database is then mined using APriori data mining with the purpose of finding the most frequently occurring patterns.

2.1. APriori Data Mining

In order to identify the frequently occurring patterns of corner features, a version of association rule data mining called APriori [1] is used. This paper includes a brief introduction to the data mining APriori algorithm, but for a more detailed explanation see [5]. The algorithm, searches the transaction database and identifies the encoded corner feature elements that co-occur most frequently within the pedestrian walking transactions with respect to negative non-walking transactions.

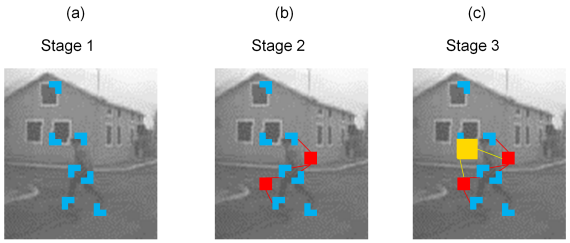


Figure 1: Example of Corners on the Dataset.

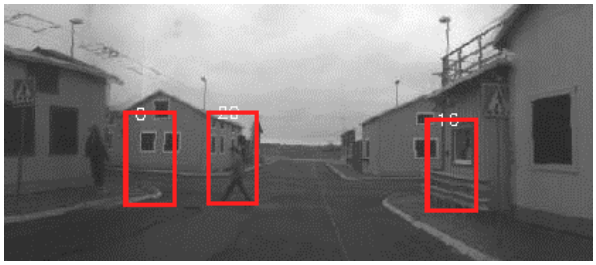


Figure 2: True/false positive detections of pedestrians.

An association rule of the form $A \Rightarrow B$ is evaluated by examining the relative frequency of its antecedent and consequent parts i.e., the set elements A and B, where A and B are encoded corner feature elements. The support for a set element is the probability that a Transaction contains the set element, i.e., $P(A, B)$. While the confidence is the conditional probability $P(B|A)$. The aim is to identify the features that are discriminative with respect to the negative set, therefore the transactions are appended with a label, ϑ, that identifies if the set is a walking or negative example. The results of data mining then include rules of the form $(A, B) \Rightarrow \vartheta$ and an estimate of $P(\vartheta|A, B)$ is given by the confidence of the rule. As the Transaction database contains both positive and negative training examples $P(\vartheta|A, B)$ will be large only if (A, B) occurs frequently in the positive examples but infrequently in the negative examples. If (A, B) occurs frequently in both positive and negative examples, then $P(\vartheta|A, B)$ will remain small and the rule ignored. A rule of corner features is distinctive if the confidence threshold is greater than 80%.

2.2. Iterative Grouping

The resulting rules from the mining will be the descriptive, distinctive compounds of corners, called frequent itemsets. These then become the basic features for the next level of mining and are then grouped within an enlarged spatio-temporal neighbourhood to form a new Transaction database, on which data mining (searching for frequently occurring sub strings) can again be performed. The process is iterated, with the final stage frequent itemsets becoming the pedestrian feature model. Fig. 1 gives an example of the grouping of corner features over 3 iteration levels. With initial 2D detected corner features in Fig. 1(a), including false positive corners detected on the building, these are ignored at the later iteration levels in Fig. 1(b and c).

For classification of unseen data, the process is identical, apart from the final iterative loop, where compound features are compared to the model learned in the training phase. A voting mechanism is used to score detected itemsets against learned/mined models. A pixel-based likelihood image for each action can be accumulated, based on the correlation between the mined trained class feature model and the detected itemsets. This provides the input for a sliding window to be applied to the image to provide the final vision based detections. This process is effective however there are often false positive detections present as shown in Fig. 2, therefore in order to further reduce the false positives, a top down logic reasoning technique is applied to the detections.

3. Logic Reasoning In Computer Vision

First-order logic (FOL) rules (represented by logic clauses) have been successfully used in computer vision to reason about propositions. Facts (represented by logic predicates) are usually the outputs of the low level visual detectors onto which logic rules are applied [10]. Feature predicates from visual detectors tend to be noisy, therefore standard *crisp* FOL fails to model the implicit stochasticity within the input data. The introduction of Fuzzy Logic into logic programming, i.e., a fuzzy extension of standard *Prolog* is therefore useful in an environment with uncertain detector inputs. Fuzzy logic is applicable to fuzzy sets, i.e., sets for which there are *degrees of membership*. This is usually formulated in terms of a membership function valued in the real unit interval $[0, 1]$. Various fuzzy logics are possible within this framework; membership functions (and therefore *truth values*) can be single values, intervals or sets of intervals within the unit interval $[0, 1]$. Fuzzy logic programming is well suited to implement methodologies comprising reasoning with uncertainty [9].

We utilise a fuzzy FOL programming platform (*Ciao Prolog*) for modelling interval-valued fuzzy logic. FOL resolution (similar to the Prolog inference mechanism) in *Ciao Prolog* but incorporating uncertainty is made possible via CLP(R) (*Constraint Logic Programming*) [6]. The propagation of truth values through logic rules is carried by means of *aggregation operators*, which subsume conjunctive operators (T-norms; min, prod etc.,) and disjunctive operators (T-conorms; max, sum etc.,) as well as hybrid operators (combinations of the previous operators) [7]. We use a constrained form of $CLP(R)$ in which clausal conjunction is formed via the *aggregation operator* 'Product' T-norm of atomic predicates where the output is a unitary interval truth value

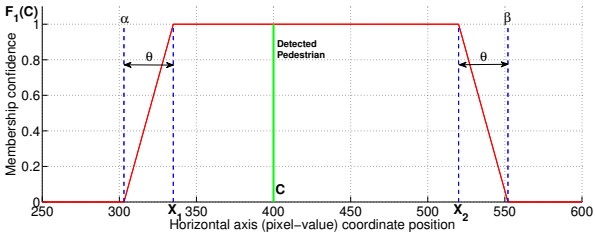


Figure 3: Fuzzy predicate '*ped_mask/2*' associates a fuzzy confidence on the basis of spatial consistency.

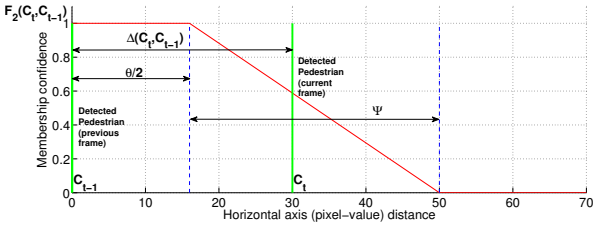


Figure 4: Fuzzy predicate '*ped_det/2*' associates a fuzzy confidence on the basis of temporal consistency.

for each predicate, in the manner of [4].

4. Fuzzy Reasoning Module

The low-level pedestrian detector (discussed in Section 2) processes individual frames in order to provide sparse feature predicates to fuzzy logic module comprising coordinate positions of bounding boxes encompassing the detected pedestrians.

We use a program module based on first-order fuzzy logic reasoning. A set of FOL clauses (*rules*) enables the system to explicitly assign a *degree-of-truth* (fuzzy confidence) to the existence of a pedestrian read from the low-level visual detector, on the basis of global spatio-temporal based logical consistency via first-order logical resolution of grounded predicates. Two principle fuzzy predicates with piecewise linear continuous membership functions are used (i.e., *ped_mask/2* and *ped_det/2*). Predicate *ped_mask/2 fuzzifies* the crisp predicates (i.e., detected pedestrian bounding boxes) with a membership confidence value relative to its coordinate positions within the junction (checks for spatial consistency). Given that $\alpha = X_1 - \theta, \beta = X_2 + \theta$; the function is defined as (refer to Fig. 3):

$$F_1(C) = max\left(min\left(\frac{C-\alpha}{\theta}, 1, \frac{\beta - C}{\theta}\right), 0\right) \quad (1)$$

The parameter C is the (horizontal axis) coordinate of the (centroid) detected pedestrian bounding box, X_1, X_2 are the coordinates of the pedestrian-crossing region (calculated from the Euclidean distance of the vehicle from junction centre), θ is equal to the length of the detected pedestrian bounding box. Given the Cartesian coordinates (DGPS); ϕ_x, ϕ_y of the junction centre, and the vehicle; γ_x, γ_y, the Euclidean distance $\Delta(\phi, \gamma)$ is given as; $\Delta(\phi, \gamma) = \sqrt{(\gamma_x - \phi_x)^2 + (\gamma_y - \phi_y)^2}$, we find the pedestrian region coordinates X_1, X_2 per-frame as follows:

$$X_1 = X_1' - \left(\frac{\Delta(\phi,\gamma)}{\eta}\right), X_2 = X_2' + \left(\frac{\Delta(\phi,\gamma)}{\eta}\right) \quad (2)$$

X_1', X_2' are the coordinates from previous frame, and η is a scaling factor. The predicate *ped_det/2* associates a fuzzy confidence to the current detection on the basis of temporal consistency. Given that C_t is the centroid of the detected pedestrian in the current frame, and C_{t-1} represents previous detections, and $\varphi = \frac{\theta}{2} + \Psi$; the function is defined as (refer to Fig. 4):

$$F_2(C_t, C_{t-1}) = max\left(min\left(1, \frac{\varphi - \Delta(C_t, C_{t-1})}{\Psi}\right), 0\right) \quad (3)$$

where the parameter $\Delta(C_t, C_{t-1}) = \|C_t - C_{t-1}\|$, and Ψ is a fuzzy parameter in the range: $\{0, 20, 40, ..., 240\}$, with the performance measure discussed in Section 5.

Each of the fuzzy predicates, *ped_mask/2* and *ped_det/2* assigns a membership confidence (e.g., V_1 and V_2 respectively) to the detected pedestrian on the basis of its spatio-temporal based logical consistency. A fuzzy rule using the aggregation operator 'Product T-norm' (i.e., $T_{prod}(V_1, V_2) = V_1 \cdot V_2$) aggregates the membership functions of each of the two fuzzy predicates. We use the predicate *ped_truth(C,V)*, to obtain the truth value $V \in [0, 1]$ of the detected pedestrian *C*, via the fuzzy rule:

$$pedestrian\_truth(C, V) :\sim T_{prod} \; ped\_mask(C, V_1), ped\_det(C, V_2). \quad (4)$$

additional *a priori* hierarchical logical predicates are asserted as *ground facts*, with fuzzy confidence 1, (e.g., $\langle at\_junction(D1, 1)\rangle$,$\langle seen\_ped\_xing(D2, 1)\rangle$ etc.,) along with FOL clauses derived from Highway Code rules. Thus a complete first-order recursive clause structure is implicit within the fuzzy logic deductive module that performs full first-order logical resolution. A selection criterion is applied to the final set of detections associated with fuzzy confidences such that those with confidence (i.e., $V > 0.5$) are asserted as true detections, while others are blanked out.

5. Experimental Results

We applied our framework to a dataset recorded from a sensor-equipped vehicle driven across a cross-junction comprising a single pedestrian. The dataset comprises external video scene recorded via three cameras (180° panoramic view), and (20 Hz) DGPS coordinates of the experimental vehicle. It comprises a subset of the original dataset, with 14 junction navigation scenarios, constituting a total of 921 frames (per frame image size of 244 x 900, at 15fps sampling).

The pedestrian detector is trained on the training examples from the walking class from the KTH

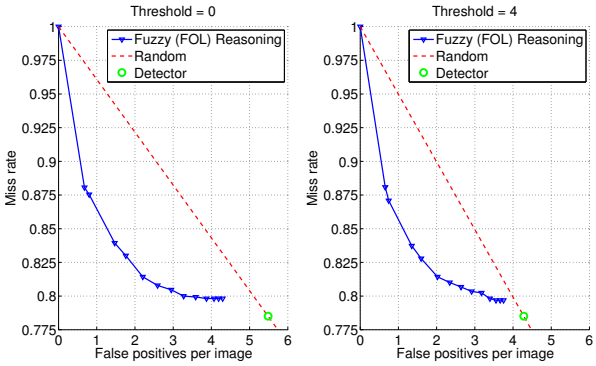


Figure 5: ROC curves for individual detector thresholds '0' (left) and '4' (right).

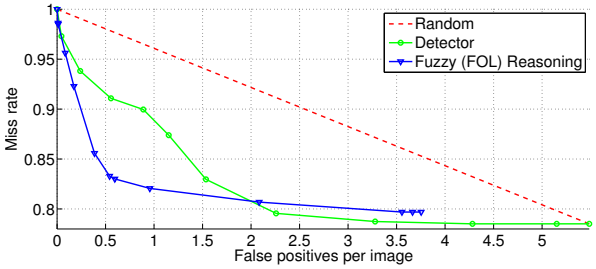
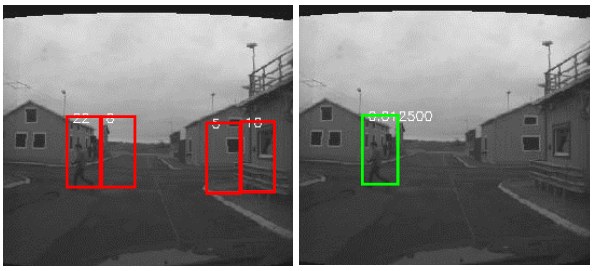


Figure 6: Convex hull of the complete set of computed ROC curves (i.e., the whole set of detector thresholds).

dataset [8], with a subset of the other classes within the KTH dataset as negative. System performance is evaluated for different detector *thresholds* and fuzzy parameter Ψ values (i.e., $\{0, 2, 4, ..., 20\}$ and $\{0, 20, 40, ..., 240\}$ respectively). A single frame evaluation (against ground-truth data) is performed on the final list of the deduced detections using the *PASCAL* measure, (i.e., the area of overlap must exceed 50%). The logic system tends to maintain a lower miss rate via logic-based consistency check, and blanking out redundant false positives as detected by the visual detector. Fig. 5 shows the ROC curves for individual detector thresholds (0 and 4). We observe that performing high level fuzzy reasoning over low-level detections improves the individual pedestrian detector performance at lower false positives per image. To measure the performance for the complete set of detector thresholds, we compute the *convex hull* of the whole set of possible ROC curves (refer to Fig. 6), and a significant increase in performance is observed, though top-down feedback does not add much to performance at very high numbers of false positives per image. Fig. 7 illustrates outputs of the system without and with top-down constraints.

6. Discussion and Conclusions

We set out a framework comprising (FOL) fuzzy reasoning for spatio-temporal logical inference of pedestrian detections from a low-level visual detector. The use of FOL theorem proving allows explicit reasoning about the existence of detected pedestrians using very sparse spatial and temporal information in the form of feature predicates from the detector. The system blanks out logically inconsistent detections by setting fuzzy constraints on low-level detector inputs. Quantitative experimental analysis illustrates improvement in performance of the system in the presence of a top-down fuzzy reasoning module. Thus top-down constraints on bottom-up low-level detections can prove to be useful in a number of different applications.



(a) (b)

Figure 7: Pedestrian detection without (a) and with (b) top-down (fuzzy FOL-based) constraints.

Acknowledgement

The work presented here was supported by EU, grant DIPLECS (FP 7 ICT project no. 215078) and EPSRC, grant ACASVA (EP/F069626/1).

References

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB'94*.
[2] N. Dalah and B. Triggs. Histograms of Oriented Gradient for Human Detection. *CVPR'05*.
[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-temporal Features. *ICCCN '05*.
[4] M. Felsberg, A. Shaukat, and D. Windridge. Online Learning in Perception-Action Systems. In *ECCV 2010 Workshop on Vision for Cognitive Tasks*, 2010.
[5] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE TPAMI*, 2011.
[6] J. Jaffar and M. J. Maher. Constraint logic programming: A survey. *J. Log. Program*, pages 503–581, 1994.
[7] S. Mu noz-Hernandez and W. Sari Wiguna. Fuzzy cognitive layer in robocupsoccer. In *IFSA '07*, 2007.
[8] C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Actions: a Local SVM Approach. *ICPR'04*.
[9] E. Y. Shapiro. Logic programs with uncertainties: a tool for implementing rule-based systems. In *8th IJCAI,'83*.
[10] V. D. Shet et al. Multivalued default logic for identity maintenance in visual surveillance. In *ECCV*, 2006.
[11] S. Walk, N. Majer, K. Schindler, and S. B. New features and insights for pedestrian detection. *CVPR'10*.