Pose Based Pedestrian Street Cross Action Recognition in Infrared Images

Raluca Didona Brehar, Cristian Cosmin Vancea, Mircea Paul Mureşan, Sergiu Nedevschi, Radu Dănescu
Technical University of Cluj-Napoca,
Computer Science Department

***Abstract*—In the context of a traffic scenario captured during night with infrared cameras we focus on pedestrian street cross action and we study the influence of the pedestrian pose with respect to the road environment on the accuracy of the action recognition model. This paper presents a complete framework that performs pedestrian cross action recognition for infrared sequences captured mainly during night but also during day time. The main contribution of the paper resides in the study of the variation in pedestrian action recognition accuracy provided by the combination of pedestrian pose-based features with several road context features given by semantic segmentation networks. The main modules of the proposed framework consist in a YOLO based infrared pedestrian detector combined with a tracking algorithm that enhances the detections. A CNN based pose estimator is applied on detected pedestrians in order to extract the relevant keypoints of the pedestrian skeleton. Several semantic segmentation networks like U-Net, FCN and PSPNet have been adapted in order to perform the semantic segmentation of the road in infrared images. Pose features are combined with road context features provided by the semantic segmentation and input to a LSTM based cross action recognition network. The obtained results provide a 90% accuracy on CROSSIR dataset.**

I. Introduction

The recognition of actions performed by pedestrians especially in traffic and focused on street cross or not cross situations can be of particular interest for autonomous driving applications, contributing to prevention and accident avoidance. Infrared sensors capture the heat emitted by objects and can be useful for night driving, in low visibility situations such as heavy rain, snow, fog, dust. In infrared images pedestrians may have a complex appearance as it can be noticed from Figure 1 due to occlusions of cold or warm objects, due to heat diffusion that makes their border silhouette blurry. Particular features like head parts, torso parts, or other body parts can be hardly distinguished in infrared images, hence approaches [1] suitable for color images in which the head or torso movement direction of the pedestrian can be relevant for his future actions, cannot be applied for infrared.

There are many solutions that perform pedestrian action recognition in the visible domain using either classical machine learning algorithms and, lately most of them exploring deep learning architectures that perform time series based action prediction [2], [3].

The presented method is focused on the recognition of the street cross action for pedestrians in monocular infrared sequences. The original aspects of the proposed method reside in:

- The fine tuning of a neural network based pedestrian pose estimator to perform pose extraction for the pedestrians detected in infrared images.
- The adaptation of several semantic segmentation networks for performing the semantic segmentation of road pixels in infrared images.
- The study and comparison of the performance in accuracy when using pedestrian pose combined with road context features derived from the semantic segmentation network.
- A generic framework for pedestrian street cross action recognition in infrared images.

The proposed system has good results for night scenes and also for day scenes achieving an overall street cross action recognition accuracy of 90% on CROSSIR dataset [2].

The rest of the paper is structured as follows: section II presents other existing approaches in the field. Section III describes the proposed framework including the network architectures with the fine tuning of parameters. The dataset used for evaluation, the parameters of the training procedure, the detailed evaluation are described in section IV, while section V concludes the paper.

Fig. 1. Samples of street cross and not cross actions in infrared images.

II. Related work

Forecasting pedestrians' intentions in advance is a challenging task in computer vision. Highly rated for automated driving safety, pedestrian crossing action anticipation methods are part of the broader family of action recognition in

978-1-6654-0976-6/21/$31.00 ©2021 IEEE

traffic scenarios, while reduced to the binary cross/not cross classification problem. The work in [1] used a Fully Convolutional Network (FCN) model of AlexNet to encode contextual information (road width, presence of traffic signals, location type) and pedestrian behavior (gait, attention), which are fed into a Support Vector Machine (SVM) classifier trained to estimate the crossing intention. The proposed Joint Attention for Autonomous Driving (JAAD) dataset used for training was enriched with relevant contextual information and behavioral tags. Unfortunately, the vast majority of such clues are not clearly distinguishable in night vision infrared scenes. From a temporal perspective, it is important to achieve a confident level of anticipation in an optimal time frame with respect to the crossing event. Some studies [4] define the crossing event immediately before the pedestrian entry on the road surface, preceded by two distinct actions: standing over the curbside and walking. Other works [5], [6] extend the typical range of pedestrian motion including bending, stopping, starting to cross and crossing-through. The number of trained classifiers is also increased, in order to deal with each scenario in particular. Although distinct from this point of view, [6] and [4] adopt a similar algorithmic approach that learns behavioral patterns from skeleton-based features analyzed on batches of consecutive frames, in which pedestrians are detected and tracked. In this context, the learning dataset (JAAD) was annotated with Time-to-Event (TTE) information for each frame: strictly positive values mark frames before the event and negative values after the event. The chosen classifier is either SVM or Random Forest (RF). They also support the hypothesis that high-level features such as skeletons are more informative than low-level features from Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) or those encoded by Convolutional Neural Networks (CNN). This is also advantageous for infrared images, which mostly preserve the appearance of the silhouette. Wang et al. [7] use AlphaPose [8] to estimate pedestrian pose from monocular sequences, while normalized positions of 9 particular keypoints from the pedestrian skeleton are selected to feed a neural network classifier with 2 hidden layers. The output of the network offers the probability of 3 behavioral states, respectively crossing, not crossing and walking along the vehicles' moving direction. Minguez et al. [9] employ Balanced Gaussian Process Dynamical Models fitted on 3D pedestrian skeleton joints, in order to incorporate spatial dynamics over time. These models facilitate prediction state and location in the future. They are capable to learn activities (e.g. standing, starting, stopping and walking) associated with specific pedestrian intentions. The skeleton model is extracted based on point clouds obtained from stereo pairs and geometrical constraints. Unfortunately, the method suffers computational bottleneck when raising the number of processed samples from input sequences. On the other hand, Rasouli et al. [10] fuse low-level and high-level contextual features in a stacked topology of Gated Recurrent Units (GRU), which demonstrate better prediction for shorter observation length. Their explanation notifies that longer observation time provides more information, however it is susceptible to more noise. Kotseruba et al. [11] advocate the importance of inferring intention information, along local visual features, pedestrian bounding-boxes and vehicles' speed, into an encoder-decoder Recurrent Neural Network (RNN) model capable to predict action. The training was performed on Pedestrian Intention Estimation (PIE) dataset [12] specifically annotated with intention scores for pedestrian instances. Yao et al. [13] train a multi-task neural network which detects the crossing intention and future actions coupled by encoder-decoder cells. The hidden states encoded within the proposed model are classified by Multi-Layer Perceptron (MLP) networks, which estimate current intent, current action and predict future actions. Inspired by neuroscience and psychological literature, the pedestrian behavior is perceived as the combined result of crossing intent (as inner will) and action in progress, hence the proposed model aims to predict both jointly.

III. Proposed framework

The proposed framework for performing pedestrian action recognition contains a Long Short Term Memory Network that based on the temporal evolution of features related to the pedestrian pose, speed, position with respect to the road, predicts the street cross or not cross action. The main components employed by the proposed framework are described in Figure 2: The infrared sequence data represents the data

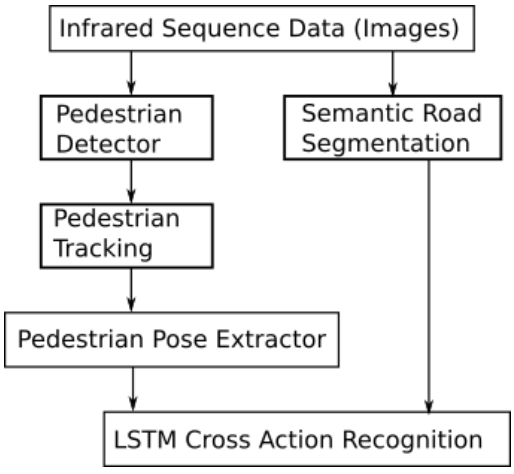


Fig. 2. Processing pipeline of the proposed pedestrian street cross action recognition framework

on which the algorithms have been trained or fine-tuned and it is formed of the sequences in the CROSSIR dataset [2]. The dataset contains 86 sequences captured with a FLIR PathFindIR camera. The sequences comprise annotated street cross and not cross actions for day and night traffic scenarios.

The pedestrian detector is based on a spatial pyramid pooling YOLO [14] type architecture that is used for determining the pedestrian detections in infrared images. It was trained and validated on the pedestrian annotations from the CROSSIR dataset. The detections are fed to the pedestrian tracking component that is composed of the following modules:

- Data association and similarity cost computation.
- Tracking selection.
- Update and refinement.

The tracking approach is a loosely coupled tracking method that follows the track by detection framework. For associating tracks and measurements, a similarity cost function based on appearance and motion has been created. The final assignment between tracks and measurements is done using an optimization algorithm. Finally, the results are refined by removing tracks which do not have any representation in the scene any more. For improving the association time between tracks and detections a measurement validation gate is used around the position of the predicted hypothesis. Only the measurements that fall within the validation gate of a track are considered in the association process for that track. The appearance score is useful in target tracking for distinguishing between different objects, even when they are close to each other, based on visual features. In our approach we have devised the appearance function to capture the textural uniqueness of each pedestrian. The proposed appearance score is presented in equation (1).

$$\begin{aligned} a(i,j) = w_{huLbp} \times huLbp(i,j) + w_{\mu s} \times \mu s(i,j) + \\ + w_{\sigma s} \times \sigma s(i,j) + w_{hs} \times hs(i,j) + \\ + w_{Ws} \times Ws(i,j) + w_{cs} \times cs(i,j) + \\ + w_{os} \times os(i,j) \end{aligned} \quad (1)$$

The appearance of the terms from the appearance score equation is the following: $huLbp(i,j)$ represents the histogram of uniform local binary pattern (LBP) in the region of interest (ROI) given by the detection, $\mu s(i,j)$ is the mean value pixel intensity distance of the ROI, $\sigma s(i,j)$ represents the variance score in the ROI, $hs(i,j)$ and $Ws(i,j)$ are the height and width distances, $os(i,j)$ represents the overlapping score and $cs(i,j)$ represents the class detection probability score. In case the similarity score based on appearance is not able to distinguish among pedestrians, we have also included a motion score that incorporates the motion pattern of the traffic participants. The expression of the motion score is defined by equation (2).

$$\begin{aligned} m(i,j) = w_{dst} * dst(i,j) + fc(i,j) + \\ + w_{\sigma m}(\sigma m(i,j)_x + \sigma m(i,j)_y) \end{aligned} \quad (2)$$

The meaning of the terms from the motion score are the following: $dst(i,j)$ represents the euclidean distance between the track position and the detection position,$fc(i,j)$ denotes the absolute difference between the flow magnitude and angle averaged in the pedestrian detection region of interest of the track and detection, and finally $\sigma m(i,j)_x$ and $\sigma m(i,j)_y$ are the deviations from the average motion pattern on the x and y directions. The weights of the appearance and motion scores have been determined experimentally. Their values are: $w_{huLbp} = 10, w_{\mu s} = 285, w_{\sigma s} = 8, w_{hs} = 10, w_{Ws} = 10, w_{cs} = 550, w_{os} = 95$, $w_{dst} = 85$, and $w_{\sigma m} = 20$.

The final similarity cost is composed by the sum of the motion and appearance costs as presented in (3).

$$\epsilon(i,j) = a(i,j) + m(i,j) \quad (3)$$

The similarity cost between each track and the corresponding detections that fall within the covariance ellipses are stored in memory and are fed as input to the Hungarian [15] algorithm. This algorithm finds the optimal track - detection assignment based on the scores. The following three scenarios can be identified after running the Hungarian optimal assignment algorithm: we can have a track matched with a detection, an unmatched detection or an unmatched track. Each of the scenarios has to be addressed independently. For an unmatched track, its position is predicted in the next frame based on the motion pattern the track has had so far. After a number of frames if no detections have been associated to a track, that track is removed. In the case of detections that are not associated, new tracks are created. For the successfully associated tracks and detections, the Kalman filter [16] is used to predict future positions and update the track state vector. A more in depth explanation of the whole data association process and track management used in this solution can be found at [2]

A regional multi-person pose estimator [17]–[19] was used by the pedestrian pose extractor (see Figure 2) in order to extract the pedestrian skeleton. The multi-pose person estimator [17] applies the skeleton feature extractor on the bounding boxes provided by the pedestrian tracking module. The infrared pedestrian bounding boxes provided by the detector and tracker are fed to a symmetric spatial transformer network and to parallel single person pose estimators as described by [17]. A pose guided proposals generator refines the obtained pose models. The default COCO key-point representation was adopted in our experiments. The results contain key-points for 17 body parts: nose, eyes, ears, shoulders, elbows, wrists, hips, knees and ankles. The (x, y) image coordinates of the points corresponding to hips, keens and ankles of a detected and tracked pedestrian in the infrared image represent part of the feature vector fed to the cross action recognition module.

The semantic road segmentation module is responsible for extracting environment context from the infrared images. It provides a road pixel mask that identifies the pixels in the image having a high probability to belong to the road. Based on them we compute the other part of the feature vector. It is represented by the average number of road pixels that are around the pedestrian legs. We consider an ellipse of width and height proportional to the pedestrian size, as shown in Figure 3 in yellow. Inside this ellipse we compute the average number of road pixels in the mask image that is output by the semantic road segmentation module.

Four semantic segmentation networks have been fined tuned and trained in order to extract the road semantic context. Fully Convolutional Network (FCN) [20], U-Net [21], Pyramid Scene Parsing Network (PSPNet) [22] and MultiNet [23] have been employed in the experiments. The four network models were previously trained on color images with several semantic classes. For the infrared semantic road segmentation

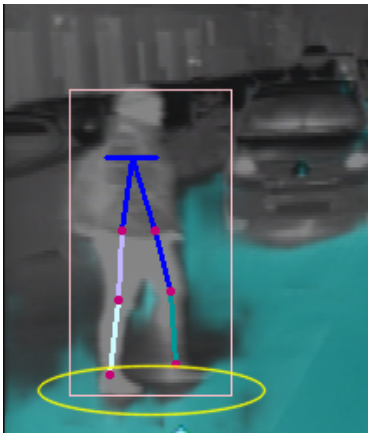


Fig. 3. Road features are formed of the average number of road pixels (marked with cyan) that reside in the ellipse marked in yellow. It shows the neighbourhood of the pedestrian legs with respect to the road pixels. The pedestrian skeleton is also depicted showing the points corresponding to shoulders, hips, knees and ankles.

the structure and parameters of each network was modified in order to accommodate two classes for semantic segmentation: background and road. The Fully Convolutional Model [20] combines the semantic information from a deep, coarse layer with appearance information provided by a fine layer and produces accurate and detailed segmentation. U-Net [21] is a classical down and up-sampling segmentation model in which only the final up-sampled feature map is utilized for segmentation. PSPNet [22] improves the segmentation accuracy of FCN type networks by using global image context for the local level predictions. It is an encoder-decoder architecture. The encoder uses dilated convolutions combined with pyramid pooling. The decoder is based on a convolution layer followed by a 8 bi-linear-upsampling operations. MultiNet [23] is also an encoder-decoder type architecture that has deep CNN as encoder and the segmentation decoder is formed based on FCN architecture.

The street cross action recognition is performed based on the time series analysis of the feature vector composed of:

- Pose features corresponding to the pedestrian legs and torso. The pose extractor has as input the pedestrian bounding boxes that result after the tracking algorithm is applied on the pedestrian detections. The (x, y) image coordinates of the following body parts are considered: shoulders, hips, knees and ankles.
- Semantic road context provided by the average number of neighbourhood pixels residing in the neighbourhood of the pedestrian legs (around the ankles).
- Pedestrian motion information which is provided by the tracking algorithm.

For recognizing the cross versus not cross action of pedestrians we have employed a Long Short Term Memory (LSTM) for time series classification. LSTM architectures have been introduced by [24] and are widely used for sequence classification problems due to the fact that their recurrent structure sustains the process of learning over time. Their structure consists of several cells, each cell being formed of gates: the input gate that controls the level of cell state update, the forget gate that controls the level of cell state reset, the cell candidate that adds information to cell state, and the output gate that controls the level of cell state added to the hidden state. We have used the LSTM implementation provided by [25]. The architecture we have employed in our experiments contains the following layers:

- The input sequence layer that has an input equal to the size of the feature vector and its output is connected to the LSTM module.
- The LSTM module layer with 20 hidden units.
- A fully connected layer with an output size equal to two which is the number of classes for the sequence classification task. The two class labels are cross and not cross.
- A Softmax layer that normalizes the values of its input data.
- A classification layer that computes the cross entropy loss.

IV. Experiments and Results

A. Data and modules setup

For training and evaluating the proposed model we have used the CROSSIR dataset [2]. The parameters we have used for each module apart are as follows:

- For training the pedestrian detector we have used YOLO [14]. The model with the highest mean average precision obtained during training for 20000 iterations is kept.
- For extracting the skeleton features we have included the above pedestrian detection model in AlphaPose [18]. Due to low memory conditions, a maximum batch size equal to two was used for the pose estimation network.
- For the segmentation networks a batch size of 4 was used. Adam algorithm is used as optimizer. The models were trained with a weight decay equal to 0.01 and a learning rate of 0.0001. Each model is trained for 100 epochs.
- The LSTM network uses a batch size equal to 32, is trained for 50 epochs using the Adam optimizer and an initial learning rate of 0.001.

B. Evaluation of Cross Action Recognition

The metrics employed for evaluating each of the modules of the proposed framework are:

- Mean average precision for the pedestrian detector.
- Mean intersection over union (mIoU) for the road segmentation networks.
- Accuracy for the street cross action recognition model.

The pedestrian detector achieves a mean average precision of 84% on the CROSSIR dataset. The skeleton points were correctly identified for 85% of the pedestrian detections. The semantic road segmentation networks attain the following mean intersection over union values:

- FCN: mIoU = 84%.
- UNet: mIoU = 85%.

- PSPNet: mIoU = 87%.
- MultiNet: mIou = 84%.

The action recognition accuracy with respect to various series length, and considering road features computed based on the four road segmentation networks is presented in Figure 4. A

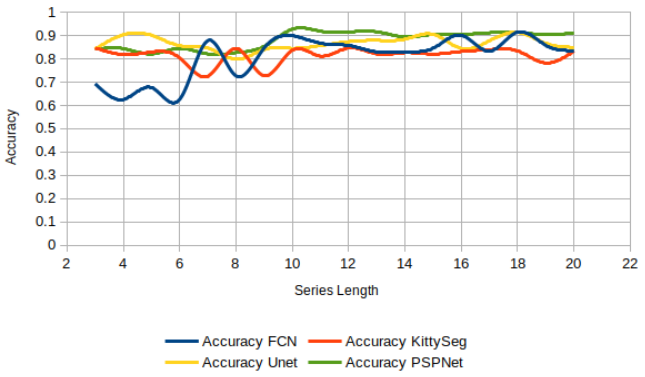


Fig. 4. Accuracy of cross action recognition with respect to series length and different road segmentation networks.

series length is defined as the minimum number of consecutive frames for which the feature vectors are computed and input to the LSTM classification network. It can be noted from Figure 4 that for a series length of four up to ten frames the accuracy varies between 0.7 up to 0.84 while for series length higher than 10 frames the accuracy is above 0.8 up to 0.9 (the maximum value being 1). This means the model can recognize the street cross action with a high probability for feature vectors that are computed along at least 10 frames before the actual cross action happens. The results obtained using skeleton features complete the previous results obtained by [2] that do not use skeleton points in their proposed model.

In what regards the type of segmentation, it can be noticed that PSPNet provides the best results, due to the high intersection over union score obtained for this type of semantic segmentation network. Nevertheless, the accuracy for all of the four road segmentation networks considered in this study is above 0.7 for series length up to 10 frames and above 0.8 for series length up to 10 frames with a maximum accuracy of 0.9.

The described framework has been tested on a system having the following parameters: i7 Processor, 16GB memory, 2080Ti GPU. Using this setup the overall inference time of the proposed framework is about 40fps (an average of 25ms per frame).

C. Demonstrative results

Some sample street cross action recognition results are shown in Figure 5. Two cross action scenarios are shown in Figure 5. In pictures (a,b,c,d,e) there is a start to cross scenario in which a pedestrian approaches the road and starts to cross. It can be noted that the cross action is recognized immediately after the pedestrians heads its legs towards the street. First two frames (a and b) report a not cross action, while the consecutive frames (c, d, e) report a cross action.

In pictures (f,g,h,i,j) a continuous cross scenario is captured. Pedestrians enter the road and continuously cross the street. The cross action is marked as a red bounding box around the detected pedestrian.

Demonstrative not cross action recognition results are shown in Figure 6. The pictures (a,b,c,d,e) in Figure 6 show a scenario where the pedestrian approaches the road and stops, without the intention of crossing. In frames (f,g,h,i,j) of Figure 6 a situation where pedestrians walk along the street is captured. The not cross action is marked by a green bounding box around the pedestrian.

V. Conclusion

The paper presents a framework for recognizing pedestrian street cross action in infrared images. It is based on a long-short term memory network that considers the temporal evolution of several pedestrian features such as motion speed and direction of movement, position of the body parts with respect to the road and in the scene. Several semantic segmentation networks have been employed in the road segmentation task in order to study and compare the accuracy of the cross action recognition with respect to the results of the segmentation network. The best model achieves a 90% accuracy of the street cross action on a benchmark infrared dataset.

Acknowledgment

This work was partly supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P4-ID-PCE-2020-1700, within PNCDI III and partly supported by the CLOUDUT Project, cofunded by the European Fund of Regional Development through the Competitiveness Operational Programme 2014-2020, contract no. 235/2020.

References

[1] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017.

[2] Raluca Didona Brehar, Mircea Paul Muresan, Tiberiu Marița, Cristian-Cosmin Vancea, Mihai Negru, and Sergiu Nedevschi. Pedestrian street-cross action recognition in monocular far infrared sequences. *IEEE Access*, 9:74302–74324, 2021.

[3] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019.

[4] Zhijie Fang and Antonio M. López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *CoRR*, abs/1910.03858, 2019.

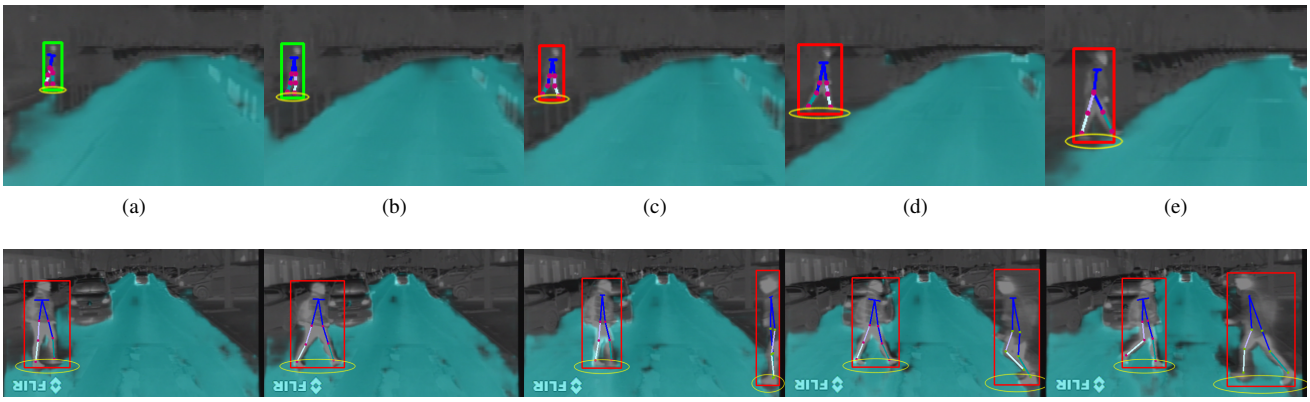
[5] Nicolas Schneider and Dariu M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In Joachim Weickert, Matthias Hein, and Bernt Schiele, editors, *Pattern Recognition*, pages 174–183, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[6] Zhijie Fang, David Vázquez, and Antonio M. López. On-board detection of pedestrian intentions. *Sensors*, 17(10), 2017.

[7] Zixing Wang and Nikolaos Papanikolopoulos. Estimating pedestrian crossing states based on single 2d body pose. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2205–2210, 2020.

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation, 2018.

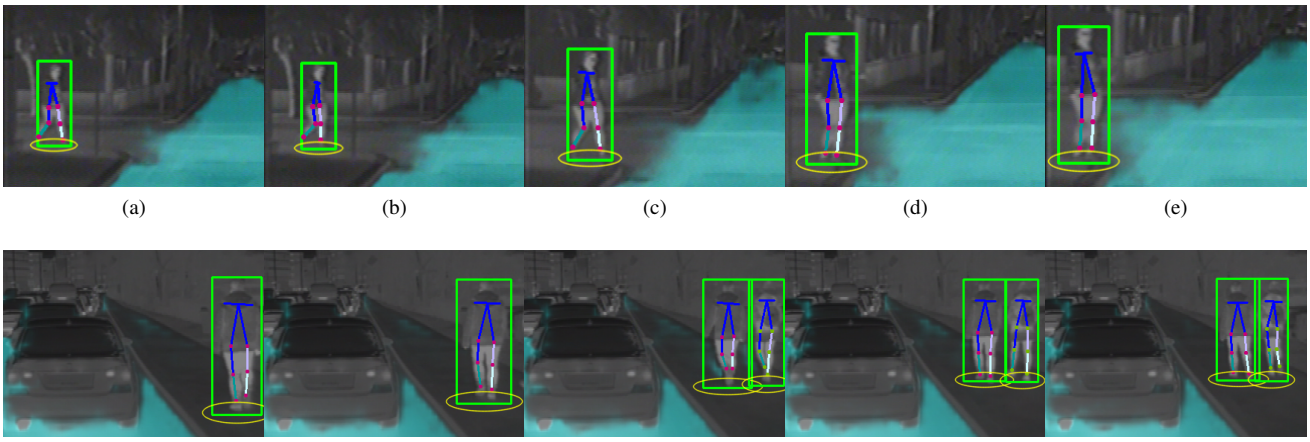
[9] Raul Quintero, Ignacio Parra, David Fernandez Llorca, and Miguel Angel Sotelo. Pedestrian path, pose and intention prediction through gaussian process dynamical models and pedestrian activity recognition, 2020.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 5. Street cross action recognition results: detected pedestrians and their skeleton points are depicted. The bounding box around the pedestrian body is Red if the street cross action is recognized.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 6. Not cross action recognition results: detected pedestrians and their skeleton points are depicted. The bounding box around the pedestrian body is Green if the not cross action is recognized.

[10] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019.

[11] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693, 2020.

[12] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6261–6270, 2019.

[13] Yu Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and Xiaoxiao Du. Coupling intent and action for pedestrian crossing behavior prediction. *ArXiv*, abs/2105.04133, 2021.

[14] Zhanchao Huang and Jianlin Wang. DC-SPP-YOLO: dense connection and spatial pyramid pooling based YOLO for object detection. *CoRR*, abs/1903.08589, 2019.

[15] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[16] Mircea Paul Muresan and Sergiu Nedevschi. Multi-object tracking of 3d cuboids using aggregated features. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 11–18, 2019.

[17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[18] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018.

[19] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[20] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2016.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[22] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[23] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*, 2016.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[25] MATLAB. *version R2020b*. The MathWorks Inc., Natick, Massachusetts, 2020.