Universitatea Babeş-Bolyai
Facultatea de Matematică şi Informatică, Departamentul de Informatică
Institut National des Sciences Appliquees de Rouen
Laboratoire d'Informatique de Traitement de l'Information et des Systemes

Multi-Task Cross-Modality Deep Learning for Pedestrian Risk Estimation

PhD Thesis

PhD student:
Eng. Dănuţ Ovidiu POP

PhD Directors:

Dr. Horia F POP	Professor	UBB Cluj-Napoca
Dr. Abdelaziz BENSRHAIR	Professor	INSA de Rouen

PhD Advisors:

Dr. Alexandrina ROGOZAN	Associate Professor	INSA de Rouen
Dr. Fawzi NASHASHIBI	Doctor HDR	INRIA Paris

8 November 2019

Contents

Keywords

- Pedestrian Detection;
- Action Recognition;
- Action Prediction;
- Pedestrian Classification;
- Cross-Modality Learning;
- Deep Learning;
- Convolutional Neural Network

Contents

Introduction

This Ph.D. thesis is the result of my research work in the intelligent transportation field to solve the problem of developing a multi-task pedestrian protection system (PPS) including not only pedestrian classification, detection and tracking, but also pedestrian action-unit classification and prediction, and finally pedestrian risk estimation.

The goal of our research is to develop an intelligent pedestrian protection component based only on a single stereo vision system using an optimal cross-modality deep learning architecture.

The system has to be able not only to detect all the pedestrians with high precision but also to track all the pedestrian paths, to classify the current pedestrian action and to predict their next actions and finally to estimate the pedestrian risk by the time to cross for each pedestrian.

So to do that we

1. First, we investigate the classification component where we analyzed how learning representations from one modality would enable recognition for other modalitie(s) in various deep learning approaches, which is termed cross-modality learning.

   The late fusion scheme connected with CNN learning is deeply investigated for pedestrian recognition based on the Daimler stereo vision dataset. Thus, an independent CNN for each imaging modality (Intensity, Depth, and Optical Flow) is used before the fusion of the CNN's probabilistic output scores with a Multi-Layer Perceptron which provides the recognition decision.

   We propose four different learning patterns based on Cross-Modality deep learning of Convolutional Neural Networks:

   (a) a Particular Cross-Modality Learning;

   (b) a Separate Cross-Modality Learning;

   (c) a Correlated Cross-Modality Learning;

   (d) an Incremental Cross-Modality Learning model.

   Moreover, we also design a new CNN architecture, called LeNet+, which improves the classification performance, not only for each modality classifier, but also for the multi-modality late-fusion scheme. Finally, we propose to learn the LeNet+ model with the incremental cross-modality approach using optimal learning settings, obtained with a K-fold Cross Validation pattern.

   This method outperforms the state-of-the-art classifier provided with Daimler datasets on both non-occluded and partially-occluded pedestrian tasks.

2. Second, we study how cross modality learning improves an end-to-end pedestrian action detection.

   We study how cross modality learning improves an end-to-end pedestrian action detection. we focus on both pedestrian detection and pedestrian action recognition based on the Joint Attention for Autonomous Driving (JAAD) dataset, applying deep learning approaches.

   The main objective of this approach is to find out if a pedestrian is crossing, or whether the pedestrian's action does not present a critical situation. The most crucial case for the pedestrian and drivers is when the pedestrian is crossing the street in the front of the vehicle, and the car cannot stop or avoid it on time.

   We introduce a unified pedestrian detection component based on deep learning, that also recognizes different pedestrian actions; this is in contrast to usual pedestrian detection methods, which only discriminate between pedestrians and non-pedestrians among other road users.

   We define four main pedestrian actions in order to find out if the pedestrian's action presents a risky situation:

   (a) the pedestrian is preparing to cross the street;
   (b) the pedestrian is crossing the street;
   (c) the pedestrian is about to cross the street;
   (d) the pedestrian's intention is ambiguous.

3. Third, we analyze the pedestrian action prediction and the estimation of time to cross. The pedestrian detection system is one of the vital components of the advanced driver assistance system because it contributes to road safety. The security of the traffic participant could be significantly improved if this system could recognize and predict pedestrian actions or even estimate the time to cross for each pedestrian.

   In this chapter, we focus on pedestrian action prediction, and estimate the time to crossing for each pedestrian. We based this work on the Joint Attention for Autonomous Driving (JAAD) dataset, applying deep learning approaches.

   We propose:

   (a) a prediction of pedestrian action using a recurrent deep learning network in order to predict the pedestrian's next actions on the short (T+1, T+2, T+3, T+4, T+5), medium (T+14) and long time (T+40);
   (b) an estimation of time to cross for a single and multiple pedestrians using recurrent deep learning network.

   We use an Long Short-Term Memory (LSTM) to estimate the pedestrian intention action using the previous 5, 14 and respectively 40 frames as time steps. We show that integrating multiple pedestrian tags for the detection part, merged with LSTM, can achieve a significant performance.

Chapter 1

Summary

Pedestrian detection is a highly debated issue in the scientific community due to its major importance for a large number of applications, especially in the fields of automotive safety, robotics and surveillance. In spite of the widely varying methods developed in recent years, pedestrian detection is still an open challenge whose accuracy and robustness has to be improved.

A pedestrian detection system has three main components: the sensors used to capture the visual data, the modality- image processing components and the classification components. In general, all these components are processed and developed together to obtain a high detection performance, but sometimes each element could be investigated separately according to the target application.

In **Chapter 1** is concerned with improving the classification task, which is the central part of the pedestrian detector.

The late fusion scheme connected with CNN learning is deeply investigated for pedestrian recognition based on the Daimler stereo vision dataset. Thus, an independent CNN for each imaging modality (Intensity, Depth, and Optical Flow) is used before the fusion of the CNN's probabilistic output scores with a Multi-Layer Perceptron which provides the recognition decision.

To achieve this aim, we develop the following methodology based on four CNNs:

1. Lenet [LBBH98] as it is a straightforward and small architecture which allows better running even on a CPU (using small image size, the default is 32x32 pixel);

2. Lenet+ which is a variation of Lenet and improves the classification performance for each modality classifier used;

3. AlexNet for its incontestable impact on machine learning due to a good balance between performance and compact architecture;

4. VGG-16 [SZ14] because of is high performance obtained with vast a architecture commonly used in pedestrian detection.

To do so, we followed the procedure below, relying on a deep learning approach:

- Investigating the performances of AlexNet and LeNet on the Caltech dataset where pedestrian bounding boxes (BBs) are more than 50 px. All BB were resized to quadratic size (64 x 64 px);

- Combining three image modalities (intensity, depth and optical flow) to feed a unique Convolutional Neural Network (CNN), using an Early fusion method and fusing the results of three independent CNNs, a using Late fusion method;

- Evaluating the LeNet architecture with various learning algorithms and learning rate policies using the classical learning method;

- a Particular Cross-Modality learning method where a CNN is trained and validated on the same image modality, but tested on a different one;

- a Separate Cross-Modality learning method which uses a different image modality for training than for validation;

- a Correlated Cross-Modality learning method where a unique CNN is learnt (trained and validated) with Intensity, Depth and respectively Optical Flow images for each frame;

- an Incremental Cross-Modality learning where a CNN is learnt with the first images modality frames, then a second CNN, initialized by transfer learning on the first CNN, is learnt on the second image modality frames, and finally a third CNN initialized on the second CNN, is learnt on the last image modality frames;

- improving the incremental cross-modality learning due to a new CNN (Lenet+) architecture that we proposed together with K-fold Cross-Validation of both the learning rate and epoch numbers;

- Learning on AlexNet and VGG-16 using the default CNNs setting with the classical learning method and respectively with the incremental cross modality deep learning method;

- Optimizing the CNNs hyper-parameters (convolution stride, kernel size, convolution number of outputs, weights of the fully connected layers) for the classical learning method and for the incremental cross modality deep learning method respectively;

- Implementing the late fusion scheme with Multi-Layer Perceptron (MLP) for both classical and incremental methods considered above.

We benchmark different learning algorithms and rate policies using the LeNet architecture. We show that the late-fusion classifier outperforms not only all single modalities but also the early-fusion classifier.

We examine all these methods with the classical learning one where each CNN is learnt and evaluated on the same image modality. We also compare all these learning patterns with the classical learning approaches within the MoE framework proposed in [EESG10, EG11] and deep Boltzmann-Machine [OW12] for the recognition of both partially occluded and non-occluded pedestrians.

In **Chapter 2**, we focus on both pedestrian detection and pedestrian action recognition based on the Joint Attention for Autonomous Driving (JAAD) [KRT16] dataset, applying deep learning approaches.

The main object of this approach is to find out if a pedestrian is crossing, or whether the pedestrian's action does not present a critical situation. The most crucial case for the pedestrian and drivers is when the pedestrian is crossing the street in the front of the vehicle, and the car cannot stop or avoid it on time.

We introduce a unified pedestrian detection component based on deep learning, that also recognizes different pedestrian actions; this is in contrast to usual pedestrian detection methods, which only discriminate between pedestrians and non-pedestrians among other road users.

We define four main pedestrian actions in order to find out if the pedestrian's action presents a risky situation:

1. Pedestrian is Preparing to Cross the street (PPC), where the pedestrian is walking/standing, pays attention or not and changes or not behavior before crossing. In the case, the actions could be: moving, looking, standing, nod, glance, handing wave, slowing down, and finally cross the street. We take into account all the actions til the cross event as PPC class. In this case, the pedestrian is definitely cross the street after these actions.

2. Pedestrian is Crossing the street (PC), where the pedestrian is observed from the point of crossing until the pedestrian crossed the road. In this case, it is mandatory to have cross action during this event but is not mandatory to have an specific event before the cross event. There are video sequences where the pedestrians are annotated only from point of crossing the street. The pedestrian could involve even other actions like looking, hand wave, speed up, nod, slow down, glance during this event.

3. Pedestrian is About to Cross the street (PAC), where the pedestrian is about to cross and pays attention and respond according to the event. In the case, the actions could be: moving, looking, standing, nod, glance, hand wave, slow down, but they will not cross the street. The pedestrian is definitely not cross the street after these actions.

4. Pedestrian intention is Ambiguous (PA), where the pedestrian is walking/ standing, and his/her intention is ambiguous. In the case, the actions could be: moving, looking, standing, glance, speed up. We consider all the actions after the pedestrian crosses the street. In this case, the pedestrian crossed the road or other evens which does not present a risk situation.

We have examined the detection part by applying a generic object detector based on the public RetineNet [LGG+17] and Faster R CNN. We have handled the Resnet50 [HZRS15] and Inception V2 CNN architectures for the classification task with the Keras public open source implementation described in [LGG+17] and Tensorflow open source implementation described in [LGG+17]. All the training process is based on the JAAD [RKT17] dataset.

The Jaad data set offers only the RGB image modality. In order to apply the In-CML, we have to extract the Depth and Optical Flow image modality and then apply an MLP of late fusion step.

To do so, we develop the following methodology relying on a deep learning approach:

- Train all pedestrian samples with the RetinaNet [LGG+17] for pedestrian detection proposes;

- Split the pedestrian Joint Attention for Autonomous Driving (JAAD) [KRT16] dataset into four classes: pedestrian is preparing to cross the street, pedestrian is crossing the street, pedestrian is about to cross the street, and pedestrian intention is ambiguous.
- Pull out the Optical Flow and Depth from JAAD dataset.
- Train all pedestrian samples using the pedestrian actions tags mentioned above with the RetinaNet using RGB, Optical Flow and Depth motion for pedestrian detection and action classification;
- Training the Incremental Deep Learning using RetinaNet for pedestrian detection and pedestrian action recognition;
- Applying the Increment deep learning Late Fusion approach using RetinaNet.

The InCML outperformed the classical detection approach on all modalities, but its performance is statistically significant only for the RGB image modality. We observed that the performance of the InCML detector is directly proportional to the achievements of each pedestrian detection actions.

The **Chapter 3** concerns in solving the pedestrian action prediction and estimate time to crossing for multiple pedestrian applying a multi-task deep learning model for.

The time to crossing estimation of pedestrians is more challenging than predicting the pedestrian action since it requires a fine analysis of the whole scene, as well as a fine analysis of the pedestrian motion. Let us emphasize that this task is challenging even for human beings.

The difficulty in solving this problem comes from the lack of public annotated data bases. Hence, there are any public data base annotated with pedestrian time to crossing, while there are several interesting huge pedestrian detection data bases (Kitti, Caltech among others). The problem is that those data bases do not provide any pedestrian action labels. The only public data set with pedestrian action tags in urban traffic environmental is JAAD [KRT16]. Since this data set does not provide directly the annotations for pedestrian time to crossing we determinate it for each pedestrian trajectory (frame sequences). We select some cues from the JAAD [KRT16] public data set in order to solve this issue and then we made our pedestrian TTC annotation for all videos.

The conventional approach for solving the difficulty of pedestrian behavior prediction is to employ a minimum of one of the dynamic elements contributing to the perception of pedestrian behavior situations such as trajectory [HTDD18], or velocity [SG13], or to anticipate the final destination of pedestrians [RWLS18]. Moreover, to achieve a high pedestrian action and movement prediction performance, it is mandatory to take into account the temporal context information in order to help predict the pedestrian behavior.

The prediction issue is commonly grouped into two categories:

1. Collision Avoidance scenarios (short-term modelling), where the goal is to react with emergency maneuvers for objects. The prediction horizon is here max. 1-2 seconds [RK15, RRL+18].
2. Long-term modelling, where the goal is to have a more comfort driving behavior. The prediction horizon here is 2+ seconds, depending on the vehicle speed and environment [RK15].

We focus on the long and short-term prediction approach of both pedestrian position and action by using an LSTM (next frames, T+1, T+2, T+3,T+4,T+5,T+14,T+40) to take into account the temporal context information (previous frames from T-5, T-14 and T-40). The LSTM input are 2D bounding box (BB) coordinates provided by the detection component mentioned above.

Whenever applying the pedestrian detection method, the LSTM input data are the pedestrian tags (label class) and BB coordinates which anticipate the next frames following the pedestrian BB coordinates and its behavior.

The estimation of time to crossing for each pedestrian is essential for the ADAS systems since it could predict if and when there could be a risky situation.

From a machine learning point of view, TTC estimation can be considered as a regression problem, where we aim at estimating an integer or real value (whether we consider a number of frames or a time in second) for each frame of a video. As the dynamic of the signal is essential to efficiently estimate TTC, we have naturally turned toward the use of a recurrent neural network to capture context of the motion. Among recurrent models, we have chosen to use LSTMs which have shown their efficiency on many sequence analysis problems.

To predict the pedestrian time to crossing, we proposed in two approaches:

- individual estimate for each pedestrian BB sequences provided by the pedestrian detector (using only PPC samples)
- multiple estimates for all detected pedestrian (using all samples).

We emphasis that the detection and prediction components are learnt independently.

The detection step is based on RetinaNet, and it has as input the entire RGB images and returns the pedestrian corresponding bounding box and its action tag.

The prediction model is based on LSTM, and it has the 2D bounding box (BB) coordinates as input data provided by the detection component. The output consists of time to crossing for each pedestrian, and it outlines over how many frames the pedestrian crosses the road. We take into account the temporal context information for the previous frames from T-5, T-14, and T-40.

The first TTC method returns a better performance than second one, but we consider the second one the promised one because it is more realistic.

Chapter 2

Conclusion

In this thesis, we have focused on developing a multi-task pedestrian protection system (PPS) which is an essential function of Advanced Driver Assistance systems (ADAS) because it reduces traffic accidents by assisting the driver and even stopping the vehicle to prevent imminent accidents. Our PPS system includes not only pedestrian classification, detection and tracking, but also pedestrian action-unit classification and prediction and, finally, pedestrian risk estimation (time to cross). This particular issue was solved by using the original cross-modality deep learning approaches.

In Chapter 1, we introduced different learning methods based on Cross-Modality deep learning of Convolutional Neural Networks (CNNs):

In the Chapter 2, we addressed severs problems:

- We applied the incremental cross-modality deep learning on the detection method
- We found out whether a pedestrian is crossing, and whether the pedestrian's action does not present a critical situation, where we have defined four main pedestrian actions:
  1. the pedestrian is preparing to cross the street;
  2. the pedestrian is crossing the street;
  3. the pedestrian is about to cross the street;
  4. the pedestrian's intention is ambiguous;
- We introduced a unified pedestrian detection component based on incremental cross-modality deep learning, which also recognizes different pedestrian actions.

The incremental cross-modality deep learning outperformed the classical detection approach on all modalities, but its performance is statistically significant only for the RGB image modality. We noticed that the performance of the incremental cross-modality deep learning detector is directly proportional to the achievements of each detection of pedestrian actions.

We extended the pedestrian detection component using the incremental cross-modality deep learning by taking into account the temporal context in order to predict the next pedestrian action. We analyzed this issue in the third part of our research without using the incremental cross-modality deep learning.

In Chapter 3, we merged the pedestrian detection component with the pedestrian action prediction and estimation of time to cross.

We developed a prediction of pedestrian action using an estimation of time to cross for a single and multiple pedestrians using a Long Short-Term Memory (LSTM)

We used a Long Short-Term Memory (LSTM) [HS97] to estimate the pedestrian intention action using the previous 5, 14, and respectively 40 frames as time steps. We showed that integrating multiple pedestrian tags for the detection part and merging with LSTM, can achieve a significant performance.

For future work, we are planning to create an end-to-end incremental cross-modality deep leaning detector-estimation time to cross approach, which will be able to do all functionalities in one step (detection, action recognition, action prediction, estimation of time to cross). In addition, we intend to apply the incremental cross-modality deep leaning model for the classification and detection of other road objects (traffic signs and traffic lights) as well as road users (vehicles, cyclists).

Bibliography

[EESG10] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 990–997, June 2010. 4

[EG11] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, Oct 2011. 4

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 10

[HTDD18] M. Hoy, Z. Tu, K. Dang, and J. Dauwels. Learning to predict pedestrian intention via variational tracking networks. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3132–3137, Nov 2018. 6

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[KRT16] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Joint attention in autonomous driving (JAAD). *CoRR*, abs/1609.04741, 2016. 4, 6

[LBBH98] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998. 3

[LGG+17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 5

[OW12] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, June 2012. 4

[RK15] E. Rehder and H. Kloeden. Goal-directed pedestrian prediction. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 139–147, Dec 2015. 6

[RKT17] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 5

[RRL+18] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf. A literature review on the prediction of pedestrian behavior in urban scenarios. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3105–3112, Nov 2018. 6

[RWLS18] E. Rehder, F. Wirth, M. Lauer, and C. Stiller. Pedestrian prediction by planning using deep neural networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5, May 2018. 6

[SG13] 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

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3