UNIVERSITÀ di VERONA

DOCTORAL THESIS

Head Pose Estimation and Trajectory Forecasting

Author:
Irtiza HASAN

Supervisor:
Prof. Marco CRISTANI

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Doctoral School of "Natural Sciences and Engineering"

May 14, 2019

UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF

Informatica

GRADUATE SCHOOL OF

Doctoral School of "Natural Sciences and Engineering"

DOCTORAL PROGRAM IN

Faculty of Computer Science

WITH THE FINANCIAL CONTRIBUTION OF
NAME OF THE FUNDING INSTITUTION

Cycle / year (1° year of attendance): XXXI / 2015

TITLE OF DOCTORAL THESIS

Head Pose Estimation and Trajectory Forecasting

S.S.D.: INF/01-INFORMATICA

Coordinator: Prof./ssa Marco Cristani

Signature _____

Tutor: Dr./ssa Fabio Galasso

Signature _____

Tutor: Dr./ssa Alessio Del Bue

Signature _____

Doctoral Student: Dott./ssa Irtiza Hasan

Signature _____

* For the list of S.S.D. please refer to the Ministerial Decree of 4th October 2000, Attachment A *"Elenco dei Settori Scientifico – Disciplinari"* available at: http://www.miur.it/atti/2000/alladm001004_01.htm

Head Pose Estimation and Trajectory Forecasting
Irtiza Hasan
PhD Thesis
Verona, May 14, 2019
ISBN:

Declaration of Authorship

I, Irtiza HASAN, declare that this thesis titled, "Head Pose Estimation and Trajectory Forecasting
" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Irtiza Hasan

Date: 18-January-2018

Acknowledgements

Doctoral dissertation in some ways is the end of a long journey. In reality, it is really an end of a beginning. My journey started by embarking on a plane and leaving the cold, mystical and exotic Finland and landing in Milan, afterwards taking a train to Verona, Italy. No adjectives could possibly define Italy. This was the beginning of my doctoral studies and the space is too little to share all memories and name all the people whom I thankful to.

First of all, to my supervisor Prof. Marco Cristani, to say I am thankful is beyond comprehension. Over the course of my Ph.D., he went beyond his usual supervisorary role to help and guide me along the way. His presence and passion for research, is simply a source of inspiration for me. Having prolonged discussions about life in general are the most cherished memories I have with you. Staying awake all night during submissions, breakfast at Swan bar, long drives to Genoa and Munich, driving me to Qestura and usual meetings at your office will be dearly missed.

During my Ph.D., I conducted two long term visits to Munich, Germany working at computer vision labs in OSRAM. I would like to thank OSRAM for providing an amazing research facility. AT OSRAM, I had the privilege to work with Dr. Fabio Galasso. It was an unbelievable experience to work with you Fabio. Your ability to simplify and analyze problems fascinates me to this day. I had the good fortune, to have someone like you at such an early stage of my career.

I would also like to sincerely thank Dr. Luigi di Stefano and Dr. Fedrico Tombari for reviewing this thesis and helping me improve the original draft.

I could not thank enough the generous grant of European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 676455.

In Verona, my earliest friend and later collaborator Dr. Francesco Setti, thumbs up mate. I would not have been able to survive the "head winds" during my Ph.D., without your support, inside and outside the domain of research. I have learnt so much from you that it is not possible to mention all of it here. I wish I could ever become a draftsman like you.

My colleagues of the SCENEUDERLIGHT project, Dr Alessio Del Bue, whose valuable feedback and help in improving the drafts of the paper is greatly appreciated. Theodoros Tsesmelis, your company during harsh cold winter evenings in OSRAM will be missed. You have been a wonderful team man who worked selflessly to make this project successful. Your never say die attitude was the driving force in this project.

My regards to colleagues and friends in Verona and Munich who made living away from home and family a bit easier, Matteo Denitto, Pietro Lovato, Davide Conigliaro, Shahbaz Baloch, Mohammad Atique, Irina Ciortan, Marco Carletti, Maya Aghaei, Herbert Kestle, Vasileios Belagiannis, Meltem Brandlmaier and Sikandar Amin, thanks to all of you.

A warm thanks to my "man for all seasons", my dear friend Dr. Saad Ullah Akram, who introduced me to this field of computer vision. During our long walks in sub zero temperatures, he explained me the most complex problems in the simplest form. To this day, he answers my queries, which at times are stupid beyond belief. I would not have done my masters let alone my PhD with your presence.

High five for my childhood friend Khawaja Umair, for sticking with me and motivating me during good and bad times of my life.

To my elder brother Affaf Hasan, who always stood by me and all those back yard sports and fights that we had during our child hood are few of the most precious memories I have.

Just like my PhD, and everything else that I have achieved in my life, it is because of my parents. To my father Abn ul Hasan and my mother Sajida Kalsoom, even writing this I become emotional. You provided me with the greatest gift any parents could ever provide, education. You guys went way beyond your means to give me the best education and more

importantly freedom of choice to pursue my dreams. Your love and support, is perhaps my only asset in life. You guys are the best parents one could ever have. This work is dedicated to you two.

Finally, to my loving wife Dr. Seejal Ejaz. Your unconditional love and support kept me sane during all of this time. During this Ph.D., at certain times, I could not manage to give you the time you deserved and still almost everyday, after work I saw you smiling and rarely complaining, this smile kept me going in all those months. I just simply cannot thank you enough.

Contents

Declaration of Authorship **v**

Acknowledgements **vii**

1 Introduction **3**
1.1 Background and motivation . . . 3
1.2 Scope of the thesis . . . 6
1.3 Contributions . . . 6
1.4 Summary of the original articles . . . 7
1.5 Outline of the thesis . . . 7

2 Background **9**
2.1 VFOA . . . 9
2.1.1 Estimation of the VFOA in open scenarios . . . 9
2.1.2 Social motivation of the VFOA as predictive model . . . 9
2.2 Trajectory forecasting . . . 11

3 Object Detection **13**
3.1 Overview . . . 13
3.2 Faster R-CNN for General Object Detection . . . 14
3.2.1 Faster RCNN for Person Detection . . . 15

4 Head Pose Estimation **19**
4.1 Introduction . . . 19
4.2 Datasets . . . 19
4.3 Proposed Methodology . . . 20
4.3.1 Training . . . 20
4.3.2 Testing . . . 21
4.3.3 Head pose classification . . . 21
4.3.4 Head pose estimation in the wild . . . 22
4.3.5 Ablation study: head pose classification . . . 23

5 LSTM Overview **27**
5.1 Brief History of LSTM . . . 27
5.2 Social LSTM . . . 28

6 Trajectory Forecasting **31**
6.1 Related works . . . 31
6.2 Datasets . . . 32
6.3 Proposed Approaches . . . 32
6.4 Experiments . . . 34
6.4.1 Quantitative results . . . 34
6.4.2 Ablation studies . . . 35
6.4.3 Experiments with HPE . . . 36

6.4.4 Qualitative results . . . 37
6.5 Data Driven Approaches for trajectory forecasting . . . 37
6.5.1 Motivation of MX-LSTM . . . 40
6.5.2 Proposed Approach . . . 41
6.5.3 Tracklets and vislets . . . 41
6.5.4 VFoA social pooling . . . 42
6.5.5 LSTM recursion . . . 43
6.5.6 MX-LSTM optimization . . . 44
6.5.7 Experiments . . . 45
6.5.8 Implementation details . . . 46
6.5.9 Evaluation Protocol . . . 46
6.5.10 Comparison with Prior Art . . . 47
Effect of head pose estimator . . . 47
6.5.11 Ablation Study . . . 47
6.5.12 Head Pose Forecasting . . . 48
6.5.13 Time Horizon Effect . . . 49
6.5.14 Substitutes for Head Pose . . . 51
6.5.15 Qualitative Results . . . 51

7 Human-Centric Light Sensing and Estimation 53
7.1 Introduction . . . 53
7.2 Ego-light-perception . . . 53
7.2.1 People detection and head-pose estimation . . . 55
7.2.2 Spatial light estimation . . . 55
7.2.3 Gaze-gathered light modelling . . . 55
7.3 Invisible light switch evaluation . . . 56
7.3.1 Dataset overview . . . 56
7.3.2 Top-view detection and head-pose estimation . . . 58
7.3.3 Person-perceived light estimation . . . 59
7.3.4 Applications of the invisible light switch . . . 62

8 Summary and Conclusion 65

Bibliography 67

List of Figures

3.1 Example of object instance localization and recognition. Usually, an object detection framework, outputs bounding boxes along with object instance label and confidence score. Illustration take from [RHGS15b] 14
3.2 Faster R-CNN, complete diagram. Illustrations adapted from [LOWFC+18] . 15
3.3 Region Proposals Network. Images taken from [RHGS15b] 16
3.4 Complete pipeline of RPN+BF.Feature maps pooled from RPN are fed into cascaded boosted forest, for accurate pedestrian detection. Images taken from [ZLLH16] 17

4.1 Network Architecture. The figure illustrates the proposed Head Pose Classification Network (HPN). The green dotted-line represents the filtered proposals at the training time and green solid represents the pedestrian detections at testing time. 20
4.2 Qualitative results of our proposed model. Jointly detecting people and estimating their head pose. 22
4.3 Regressing the head pose of the person in a real world surveillance scenario. . 23

5.1 Illustration of the vanishing gradient problem in RNNs. One could see the effect of gradient vanishes over time (lighter color) 28
5.2 Detailed diagram of a Simple Recurrent Network unit and LSTM. Image courtsey [GSKSS17] 29
5.3 Description of Social LSTM. Authors proposed one LSTM per person. Image adapted from [AGRRF+16] 30
5.4 Description of Social LSTM. Authors proposed one LSTM per person. Image adapted from [AGRRF+16] 30

6.1 Graphical explanation on the selection of pedestrians to be taken into account for the avoidance term. The large blue dot represents the target pedestrian, the green dots are the pedestrians he/she tries not to collide to, and the small red dots are the pedestrians he/she is not aware of because out of the view frustum. (Best viewed in colors.) 32
6.2 Θ angle of the VFOA in relation with the Mean Average Displacement error . 36
6.3 Examples of predicted trajectories on UCY (first two rows) and Zara01, Zara02 (last two rows). Our proposed model is very precise in the prediction of highly non-linear trajectories, where the other approaches such as LTA [PESV09] and SF [YBOB11] are less accurate due to the fixed destination points. In particular, our method is able to easily capture short term deviations from the desired path. 38
6.4 Illustration of common failure cases of trajectory forecasting. Acceleration, deceleration and static groups are common failure cases across all approaches. 39

6.5 Motivating the MX-LSTM: a) analysis between the angle discrepancy ω between head pose and movement, the pedestrain smoothed velocity and the average errors of different approaches on the UCY sequence [LCL07]; b) correlation between movement angle β and head orientation angle α when the velocity is varying (better in color). 40
6.6 A graphical interpretation of tracklets and viselets. a) tracklets $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_{t+1}^{(i)}$ and vislet anchor point $\mathbf{a}_t^{(i)}$; b) Social pooling leveraging the Visual Frustum of Attention; c) angles for the correlation analysis. 42
6.7 VFOA pooling: For a given subject, he will try to avoid collision with the people who are inside his view frustum (blue circle). Others (red circle), will not influence his trajectory as they are no in his view frustum. 45
6.8 Qualitative results: a) MX-LSTM b) Ablation qualitative study on Individual MX-LSTM (better in color). 52

7.1 The Invisible Light Switch Framework 54
7.2 Modeling of LSC from Human Perspective 56
7.3 ILS Dataset 57
7.4 Light Management Installation 57
7.5 Light Activations in the ILS Dataset 58
7.6 Modeling Human Occupancy and Posture 58
7.7 Confusion Matrices of the Head Pose Estimator 59
7.8 ILS Illumination Estimation Error 60
7.9 Illumination Map - Scene 1 61
7.10 Illumination Map - Scene 2 61
7.11 ILS Qualitative Analysis 63

List of Tables

4.1 Comparison of head pose classification accuracy in regard to image scale variation. . . . 21
4.2 Head pose estimation in the wild. For LAMR, lower is better. . . . 24
4.3 Head pose classification accuracy on oracle. . . . 25
4.4 Pedestrian detection results. For LAMR lower is better . . . 25

6.1 Dataset Statistics . . . 32
6.2 Mean Average Displacement (MAD) error for all the methods on all the datasets. . . . 35
6.3 Final Average Displacement (FAD) after 12 frames (4.8 seconds) for all the methods on all the datasets. . . . 35
6.4 Model parameters obtained from training sequences . . . 35
6.5 Mean Average Displacement (MAD) with and without the view frustum condition in the avoidance term. . . . 35
6.6 Final Average Displacement (FAD) with and without the view frustum condition in the avoidance term. . . . 36
6.7 Mean Average Displacement (MAD) for state of the art methods with destination point estimated from the head orientation. . . . 36
6.8 Mean Average Displacement error with quantized annotated head pose and with real head pose estimator. . . . 37
6.9 Mean and Final Average Displacement errors (in meters) for all the methods on all the datasets. The first 6 columns are the comparative methods and our proposed model trained and tested with GT annotations. MX-LSTM-HPE is our model tested with the output of a real head pose estimator [HTGDC17]. The last 3 columns are variations of our approach trained and tested on GT annotations. . . . 46
6.10 Mean angular error (in degrees) for the state-of-the-art head pose estimator [HTGDC17], and our model fed with manual annotations (MX-LSTM) and estimated values (MX-LSTM-HPE). . . . 49
6.11 Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 8 frames. . . . 49
6.12 Mean Average Displacement (MAD) error when changing the observation period. Forecasting horizon is kept constant at 12 frames. . . . 50
6.13 Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 16 frames. . . . 50
6.14 MAD errors on the different datasets . . . 51

7.1 Average Estimated Illumination Error . . . 59
7.2 ILS Quantitative Analysis . . . 62

List of Abbreviations

VFOA	**V**isual **F**rusutm of **A**ttention
CNN	**C**onvolutional **N**eural **N**etworks
HPN	**H**ead **P**ose **N**etwork
RPN	**R**egion **P**roposal **N**etwork
R-CNN	**R**egion **C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong **S**hort **T**erm **M**emory
MX	**Mi**ng
BP	**B**ack**P**ropagation
RTRL	**R**eal **T**ime **R**e-current **L**earning
ILS	**I**nvisible **L**ight **S**witch

Dedicated to my parents

UNIVERSITY OF VERONA

Abstract

Faculty of Computer Science
Doctoral School of "Natural Sciences and Engineering"

Doctor of Philosophy

Head Pose Estimation and Trajectory Forecasting

by Irtiza HASAN

Human activity recognition and forecasting can be used as a primary cue for scene understanding. Acquiring details from the scene has vast applications in different fields such as computer vision, robotics and more recently smart lighting. In this work, we present the use of Visual Frustum of Attention(VFOA) for scene understanding and activity forecasting. The VFOA identifies the volume of a scene where fixations of a person may occur; it can be inferred from the head pose estimation, and it is crucial in those situations where precise gazing information cannot be retrieved, like in un-constrained indoor scenes or surveillance scenarios. Here we present a framework based on Faster RCNN, which introduces a branch in the network architecture related to the head pose estimation. The key idea is to leverage the presence of the people body to better infer the head pose, through a joint optimization process. Additionally, we enrich the Town Center dataset with head pose labels, promoting further study on this topic. Results on this novel benchmark and ablation studies on other task-specific datasets promote our idea and confirm the importance of the body cues to contextualize the head pose estimation. Secondly, we illustrate the use of VFOA in more general trajectory forecasting.. We present two approcahes 1) a handcrafted energy function based approach 2) a datat driven approach.

First, Considering social theories, we propose a prediction model for estimating future movement of pedestrians by leveraging on their head orientation. This cue, when produced by an oracle and injected in a novel socially-based energy minimization approach, allows to get state-of-the-art performances on four different forecasting benchmarks, without relying on additional information such as expected destination and desired speed, which are supposed to be know beforehand for most of the current forecasting techniques. Our approach uses the head pose estimation for two aims: 1) to define a view frustum of attention, highlighting the people a given subject is more interested about, in order to avoid collisions; 2) to give a short time estimation of what would be the desired destination point. Moreover, we show that when the head pose estimation is given by a real detector, though the performance decreases, it still remains at the level of the top score forecasting systems.

Secondly, recent approaches on trajectory forecasting use tracklets to predict the future positions of pedestrians exploiting Long Short Term Memory (LSTM) architectures. This paper shows that adding vislets, that is, short sequences of head pose estimations, allows to increase significantly the trajectory forecasting performance. We then propose to use vislets in a novel framework called MX-LSTM, capturing the interplay between tracklets and vislets thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. At the same time,MX-LSTM predicts the future head poses, increasing the standard capabilities of the long-term trajectory forecasting approaches.

Finally, we illustrate a practical application by implementing an Invisible Light Switch (ILS). Inseid ILS detection, head pose estimation and recognition of current and forecast human activities will allow an advanced occupancy detection, i.e. a control switch which turns lights on when the people are in the environment or about to enter it. Furthermore, this work joins research in smart lighting and computer vision towards the ILS, which will bring both technologies together. The result light management system will be aware of the 3D geometry, light calibration, current and forecast activity maps. The user will be allowed to up an illumination pattern and move around in the environment (e.g. through office rooms or warehouse aisles). The system will maintain the lighting (given available light sources) for the user across the scene parts and across the daylight changes. Importantly, the system will turn lights off in areas not visible by the user, therefore providing energy saving in the invisible.

Chapter 1

Introduction

Human activity recognition and forecasting can be used as a primary cue for human-centric scene understanding. Acquiring details of human activity from the scene has vast applications in different fields such as computer vision, robotics and more recently smart lighting. Beforehand knowledge of the scene gives the ability to control the lighting based on ongoing and future activities that might take place in the observed scene. In this regards, we propose the use of visual frustum of attention (VFOA) for scene understanding, activity recognition and activity forecasting. VFOA identifies the volume of a scene where fixation of a person may occur, it can be inferred from head pose estimation, and it is crucial in scenarios where precise gazing information cannot be retrieved. The VFOA can be used as a fundamental feature for activity forecasting so leading to a smart lighting system that predict the forthcoming actions and thus activate the correct lighting pattern.

1.1 Background and motivation

A fundamental step in understanding human activity is to find people presence in images. From this information then we might attempt to obtain a description of the VFOA. For this reason people (or pedestrian) detection is a preliminary stage of all the head pose estimation approaches, in fact all the methods presented so far assume that the position of the heads to be processed are either given as a ground truth information or computed by an off-the-shelf detector. Despite the correlation between these two tasks is evident, these have been never investigated as a joint problem, in which (head) detection and (head pose) classification are not distinct operations, but are different terms of a unique optimization function.

In this thesis, we propose a unified framework to address the tasks of pedestrian detection and head pose estimation as a joint problem, leveraging on deep networks and reformulating the very popular Faster R-CNN [Gir15] architecture to infer head pose in unconstrained scenarios jointly with pedestrian detection.

The head pose is an important visual cue for several computer vision applications. In surveillance videos, the joint attention of people towards a direction can signal a particular event is happening [GXH10]. In social signal processing, the head orientation is necessary to infer group formations [CBPFT+11a] and capture social roles, such as leaders/followers [Eng94]. Most recently, the head pose has been used for novel marketing strategies and architectural design, as a proxy to personal interest in goods, impact of adverts and space utilization [DLB10].

The head pose estimation (HPE) problem is challenging in particular when people are captured at far and not yet addressed "in the wild". In many practical problems, such as video surveillance, HPE input is a head region as small as a 24×24 head pixel. This information alone is not enough to obtain reliable performance in HPE [TSCM13a], and multi-view camera setting are necessary [RSRVL+14]. Recently, few works used deep learning to regress the head pose of a person such as [OCM07; MR15; LLWLP16; RPC19], the underlying idea is to use features from the convolutional layers and predict the head pose using L2-loss.

[LJMMH17], were among the fist to use a Gaussian mixture model couple with CNNs to regress the head pose.

This research proposes to increase HPE performance by leveraging information from the entire body of the person instead of using the head information only.

Specifically, we enrich the recent Faster R-CNN [RHGS15a] architecture with a branch specialized on the yaw modeling of the head pose (in this work, we focus on yaw, keeping the modeling of pitch and roll as future goals), called Head Pose Network (HPN). The idea is to jointly optimize the pedestrian detection and the HPE tasks, in order to establish and exploit a structural connection between the appearance of the body and the head pose.

The experiments, on the Town Center dataset [BR09b] and on standard benchmarks (oracle head detections are provided) show the net potentialities of our approach; additional ablation studies confirm that the body estimation, even if noisy, greatly improve the head pose estimation.

Most literature on HPE has considered high resolution images [MT09], which does not apply to surveillance videos. More recently, HPE from low resolution images [TFSMC10; OGX09; TSCM13a] has emerged to address the surveillance camera viewpoint. Here several state of the art works leverage SVM [OGX09], deep neural networks [CRSBC+15; YYJ15] and random forest [LYO15a]. Differently from this, we consider the joint HPE estimation and the person detection and we argue for the virtues of their joint training.

Our work further relates to literature on people detection, which can be widely grouped into integral channel features+boosting [DABP14], deformable part model [FGM10] and deep neural network techniques [LLSXF+15; TLWT15]. Interestingly, only recently CNN techniques have achieved the state of the art [ZLLH16] on the Caltech benchmark [DWSP09] but this dataset has images of people that differs consistently from a video surveillance scenario as the one in the Town Center scenario.

Furthermore in this thesis, we show that the head pose estimation can be used to design an effective predictive model for pedestrian path prediction, capable of boosting systematically the performance of tracking approaches at the state of the art. Starting from merely speculative investigations [Cv80], recent studies on artificial lighting in public spaces [FUCH15; FUY15], inspired from neuroscience research [PV03; MP07], analyzed the pedestrians critical visual fixations when walking on public spaces. Critical visual fixations are different from simple fixations because they entail cognitive processes focused on the object of the fixation, while simple fixations may be the effect of daydreaming or task-unrelated thoughts [FUCH15]. The goal of the research was, thanks to eye tracking portable devices, to check which are the objects that have been (critically) fixated, categorizing them in eight categories: person, path (pathway in direction of travel), latent threat, goal, vehicle, trip hazards, large objects, general environment. The results suggested that the path and other people are the more frequently critical observations, with a tendency for other people to be fixated at far distances and the path to be fixated at near distances.

Pedestrian forecasting stands for anticipating the future, based on observations and on prior understanding of the scene and actors. Further to past trajectories, forecasting the position of pedestrians requires therefore an intuition of the people goals [PESV09], their social interaction models [RSAS16; AGRRF+16; GJFSA18a], the understanding of their behavior [AMBT06; LK16; MHLK17] and possible interactions with the scene [KZBH12].

Building on this, we can assume that a robust and continue estimation of the head orientation of a pedestrian, and thus of its VFOA, would help in predicting its future close path, accounting for the other elements which are in the scene (pedestrians, obstacles). The idea is to create attentive maps of the scene (one for each pedestrian) that at each pixel do contain the probability of passing there. These maps are created by accumulating VFOAs at each time step, so that a steady head orientation would predict with higher emphasis a possible path in the future, than what a frequently changing head pose can do. In addition, head poses

of other people act on the attention maps discouraging potential trajectories that may lead to collisions.

Forecasting is important for tracking [YBOB11; SAS17; LYU18], especially in the case of missing or sparse target observations. In addition, it is a crucial compound for early action recognition [XPDY08; Ryo11; HD14] and more in general for surveillance systems [CLFK01; FMW00]. Furthermore it is indispensable for deploying autonomous vehicles, which should avoid collisions [BFS17], and for conceiving robots, respectful of the human proxemics [DRS11; Hal66; KKSB12; MHB16; TK10; ZRGMP+09].

Forecasting trajectories from images, however, is a complex problem and, probably for this reason, it has only recently emerged as a popular computer vision research topic. In particular, the modern re-visitation of Long Short Term Memory (LSTM) architectures [HS97], has enabled a leap forward in performance [SZDZ17; SDZLZ16; SYMHD17; VS17; GJFSA18a]. On one side, LSTM has allowed a seamless encoding of the social interplay among pedestrians [AGRRF+16; GJFSA18a]. On the other side, the new systems have abandoned cues demanding *oracle* knowledge, such as the person destination point [PESV09], and are therefore causal predictions.

In our work, we differ from previous approaches, because we additionally leverage the visual attention of people for forecasting, further to their position. We infer their visual attention from their head pose. We are motivated by the strong correlation between the past short-term trajectories of the people (sequences of (x, y) position coordinates, named *tracklets*) and their corresponding sequences of head pan orientations, which we name *vislets*. Our novel contribution is supported by several sociological studies [Cv80; DR12; FUCH15; FUY15; FWK11; PV03; VCDPL13]and here motivated by statistical analysis conducted on the UCY dataset [LCL07].

This work introduces MiXing LSTM (MX-LSTM), an LSTM-based framework that encodes the relation between the movement of the head and people dynamics. For example, it captures the fact that rotating the head towards a particular direction may anticipate turning and starting to walk (as in the case of a person leaving a group after a conversation). This is achieved in MX-LSTM by mixing the tracklet and vislet streams in the LSTM hidden state recursion by means of a cross-stream full covariance matrix. During the LSTM back-propagation, the covariance matrix is constrained to be positive-semidefinite by means of a log-Cholesky parameterization. This generalizes the approach of [AGRRF+16] (specific to the 2D positions *x,y* of people) to model state variables of dimensions four (position and head pose) and higher.

Vislets allow for a more informative social interplay among people. Instead of considering all pedestrians within a radius, as done in [AGRRF+16; VS17], here we only consider those individuals whom the person can see. Furthermore MX-LSTM forecasts both tracklets and vislets. Predicting visual attention in crowded scenarios makes a novel frontier for research and new applications.

We have first presented MX-LSTM in [HSTDG+18]. This paper extends our previous work in four directions: 1) we include a comprehensive evaluation of its performance on the UCY video sequences (Zara01, Zara02 and UCY) [LCL07] and on the TownCentre dataset [BR09a], following standard evaluation protocols of trajectory forecasting [PESV09; AGRRF+16; GJFSA18a]. 2) we provide an extensive evaluation with the most recent approaches to show that MX-LSTM retains overall the best performance. MX-LSTM has the ability to forecast people when they are moving slowly, the Achille's heel of all the other approaches proposed so far. Additionally, here we provide novel experiments to test its robustness by predicting in the longer-term horizon and by using an estimated (thus noisy) head pose estimator [LYO15b] also for training. 3) We verify that vislets help beyond the mere larger model capacity, by testing MX-LSTM with position-related variables replacing vislets.

4) we provide novel qualitative illustrations, detail failure cases; and finally we perform novel simulations, which uncover how the learned head poses affect the people motion.

1.2 Scope of the thesis

Head pose estimation and trajectory forecasting are two challenging problems in computer vision. In this thesis, we try to address both of the problems. Importantly, this thesis shows how these two separate problems can be addressed and furthermore how head pose estimation can be used as a pivotal cue in trajectory forecasting. For both of the problems this work points out the challenging cases. Head pose estimation gets significantly hard in crowded space due to small head size and occlusion. We propose the use of full body to overcome the aforementioned problems for head pose estimation. We analyzed that for trajectory forecasting almost all approaches fail when the velocity of the person becomes low due unpredictable behaviour. This Achilles heel of trajectory forecasting can be handled more efficiently by taking into account the head movements since persons walking trajectory and head poses are generally correlated. The methods proposed in this thesis are tested on public benchmarks and compare against state of the art. This thesis also discusses unsolved cases of trajectory forecasting, especially at low velocities.

1.3 Contributions

The first contribution of this thesis is a Head Pose Network (HPN). The HPN tries to estimate person's head orientation. Additionally, this thesis for the first time discusses the fact that it is beneficially to estimate head pose through the whole body. Since in crowded scenarios, head is often due to its tiny size is partially occluded and head pose has subtle differences between different viewing angles. This approach also partially addresses the occlusion part and whole body in crowded scenarios in most cases is more illustrated.

Additionally, we explore the correlation between people trajectories and their head orientations. We argue that people trajectory and head pose forecasting can be modelled as a joint problem. Recent approaches on trajectory forecasting leverage short-term trajectories (aka tracklets) of pedestrians to predict their future paths. In addition, sociological cues, such as expected destination or pedestrian interaction, are often combined with tracklets. In this paper, we propose MiXing-LSTM (MX-LSTM) to capture the interplay between positions and head orientations (vislets) thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. We additionally exploit the head orientations as a proxy for the visual attention, when modeling social interactions. MX-LSTM predicts future pedestrians location and head pose, increasing the standard capabilities of the current approaches on long-term trajectory forecasting. Compared to the state-of-the-art, our approach shows better performances on an extensive set of public benchmarks. MX-LSTM is particularly effective when people move slowly, i.e. the most challenging scenario for all other models. The proposed approach also allows for accurate predictions on a longer time horizon.

Finally, this work proposes an application in the domain of smart lighting. Where we combine novel research in computer vision and smart lighting. Chapter 7 is the joint work combined with Mr. Theodoros Tsesmelis and this chapter is overlapping between his thesis and mine.

1.4 Summary of the original articles

This thesis is based on five articles. In the articles contributions mentioned above are discussed and explained.

In Paper 1, we enriched a state of the art object detector Faster R-CNN [RHGS15b] with a head pose network(HPN). The key idea is to leverage the presence of the people body to better infer the head pose, through a joint optimization process.

Paper 2, In this paper we show the importance of the head pose estimation in the task of trajectory forecasting. This cue, when produced by an oracle and injected in a novel socially-based energy minimization approach, allows to get state-of-the-art performances on four different forecasting benchmarks, without relying on additional information such as expected destination and desired speed, which are supposed to be know beforehand for most of the current forecasting techniques. Our approach uses the head pose estimation for two aims: 1) to define a view frustum of attention, highlighting the people a given subject is more interested about, in order to avoid collisions; 2) to give a short time estimation of what would be the desired destination point.

Paper 3, discusses recent approaches on trajectory forecasting, use of tracklets to predict the future positions of pedestrians exploiting Long Short Term Memory (LSTM) architectures. This paper shows that adding vislets, that is, short sequences of head pose estimations, allows to increase significantly the trajectory forecasting performance. We then propose to use vislets in a novel framework called MX-LSTM, capturing the interplay between tracklets and vislets thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. At the same time, MX-LSTM predicts the future head poses, increasing the standard capabilities of the long-term trajectory forecasting approaches. With standard head pose estimators and an attentional-based social pooling, MX-LSTM scores the new trajectory forecasting state-of-the-art in all the considered datasets.

Paper 4 extends paper 3. In this paper we analyze the correlation between people trajectories and their head orientations, and we argue that forecasting can benefit from the joint optimization of these two features. The proposed approach, MiXing-LSTM (MX-LSTM), is a novel framework able to capture the interplay between positions and head orientations thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. Furthermore, we exploit head orientations as a proxy for the visual attention when modeling social interactions. Compared to the state-of-the-art, our approach shows better performances on an extensive set of public benchmarks, when compared with the best performing competitors, and is proved to be particularly effective when people moves slowly, *i.e.* the most challenging scenario for all the other models in the literature. The proposed approach also allows for accurate predictions on a longer time horizon. Besides the future locations, MX-LSTM additionally predicts future head poses, increasing the standard capabilities of the long-term trajectory forecasting approaches.

Finally, in paper 5 we combine the research proposed in paper 1 to 4 to suggest a practical application in the field of smart lighting. This work is shared with the first author of the paper. My contribution to this work is regarding the human perception part, where we are required to detect and estimate the head pose of the person. Experiments and the drafting of the relevant subsections (detection and head pose estimation) are primarily done by me.

1.5 Outline of the thesis

The rest of the thesis is organized as following. In chapter 2 we introduce the computer vision field, some of the topics include deep learning, object detection, head pose estimation and trajectory forecasting.

Chapter 3, discusses the brief history of object detection and furthermore it brings into light Faster R-CNN [RHGS15b]. This chapter also discusses common problems that Faster R-CNN faces when applied to pedestrian detection.

Chapter 4, discusses the HPN, joint optimization of HPN along with person detection and finally comparison with state of the art.

In Chapter 5, we present LSTMs. Additionally we discuss a recent state of the art approach in trajectory forecasting Social LSTM [AGRRF+16].

In chapter 6, we discuss the field of trajectory forecasting. We describe the early energy based approaches and discuss their strength and weaknesses. Subsequently, we discuss the recently proposed data driven approaches and finally we describe the proposed MX-LSTM, and how ti performs w.r.t the state of the art.

Chapter 7 a joint work on a practical application in smart lighting name Invisible Light Switch.

Chapter 8, concludes the thesis. It discusses the current underlying issues as well as limitations of the proposed work. Additionally, this chapter also sheds light on potential future directions to explore.

Chapter 2

Background

2.1 VFOA

2.1.1 Estimation of the VFOA in open scenarios

In this section we review those approaches that employ the VFOA in unconstrained scenarios, with no high resolution sensors to capture the precise gazing activity. The earlier works that focus on estimating VFOA on low resolution images were [SFYW99; RR06b] and [BO04], jointly with the pose of the person. VFOA has been used primarily for spotting social interactions: in [BCTFP+13] the head direction serves to infer a 3D visual frustum as approximation of the VFOA of a person. Given the VFOA and proximity information, interactions are estimated: the idea is that close-by people whose view frustum is intersecting are in some way interacting. The same idea has been explored, independently, in [RR11]. In [SBOG08], the VFOA was defined as a vector pointing to the focus of attention, thanks to an approximate estimation of the gazing direction at a low resolution; in that work the goal was to analyze the gazing behavior of people in front of a shop window. The projection of the VFOA on the floor was modeled as a Gaussian distribution of "samples of attention" ahead of a pedestrian in [CBPFT+11b]: the higher the density, the stronger the probability that in that area the eyes' fixation would be present. More physiologically grounded was the modeling of [VMCHP+16]: in that work, the VFOA is characterized by a direction θ (which is the person's head orientation), an aperture $\alpha = 160^\circ$ and a length l. The latter parameter corresponds to the variance of the Gaussian distribution centered around the location of a person. Even in this case, samples of attention were used to measure the probability of a fixation: a denser sampling was carried at locations closer to the person, decreasing in density in zones further away. The frustum is generated by drawing samples from the above Gaussian kernel and keeping only those that fall within the cone determined by the angle α. In [ZH16], the aperture of the cone can be modulated in order to mimic more or less focused attention areas.

In all these approaches, VFOA has been employed to capture group formations. At the best of our knowledge, this is the first work where the VFOA is employed for the estimation of a predictive model.

2.1.2 Social motivation of the VFOA as predictive model

In this section we motivate the usage of the VFOA as predictive model in a context of tracking, taking from the sociological literature. One of the earlier interesting studies was [Cv80], investigating the most critical visual tasks that pedestrians have to perform while wandering; it suggested that these tasks are obstacle detection, facial recognition of other pedestrians and visual orientation, but these assumptions have not been validated nor have been weighted for relative importance. Eye tracking was thus adopted to get quantitative results, firstly on controlled laboratory settings. In [PV03], participants walk three 10m paths; two of the paths have regularly- or irregularly-spaced footprints that subjects have to step on, the third path has no footprints. The results showed that for the 59% of total fixation time, gaze was held

on the near path at a fixed distance slightly ahead of the pedestrian, with fixations on the footprints accounting for 16%. The relationship between speed and width of the VFOA was investigated in [VCDPL13], where cyclists were asked to ride a 15m path in an internal environment with three lane widths and at three different speeds. Result showed that narrower path and higher speed demand a more restricted visual search pattern and fewer task-irrelevant fixations.

Despite the expected results, these studies have been criticized of being unnatural, taking place in constrained scenarios that lack the distracting features that would be present in the real world, such as other pedestrians, buildings and eye-catching objects. Mobile eye-tracking systems have solved this problem, allowing eye-tracking to be carried out in ecological outdoor situations. The first studies of this kind showed that 21% of fixation time was directed towards people, 37% towards the path, and 37% towards other objects [FWK11], with the percentage of fixations toward the path augmenting during the night hours (40-50%) [DR12].

Even these results were criticized: in facts, the object or area that a person fixates does not always reflect where her attention is focused, due for example to daydreaming activities or task-unrelated thoughts [DR12; FFK13; FUCH15]. Alternative protocols were studied, for example focusing on shifts in fixations, which should reflect changes in where our attention is focused [FG03]; unfortunately, the connection between eye movements and attention is still subject of studies. For this reason, in [FUCH15; FUY15] the concept of *critical fixation* was exploited: critical visual fixations are different from simple fixations because they entail cognitive processes focused on the object of the fixation. The way to detect critical fixations is based on the presence of a secondary task: other than the primary task (walking in an environment), a secondary task has to be carried out (pressing a button after having heard an auditory stimulus). A delay in the completion of the secondary task is used to identify critical fixations. In the study of [FUCH15], participants were asked to walk a short (900m) and heterogeneous route (road crossings, uneven terrain, residential areas and crowded plazas) whilst wearing the eye tracking equipment and carrying out the dual task. Critical fixations were categorized in eight categories: *person, path* (pathway in direction of travel), *latent threat, goal, vehicle, trip hazards, large objects, general environment*. Results showed that the more frequently critical observations are on the path (22%), the people (19%) and the goal (15%) with a tendency for other people to be fixated at far distances ($> 4m$) and the path to be fixated at near distances ($\leq 4m$). In addition, it is postulated that fixations at people are due to the need of perceive their motion (speed and direction) [FUY15].

These results motivated us to exploit the VFOA for collecting plausible locations of fixations (not precisely estimable in a surveillance scenario where the camera is far from the people). In particular, we consider physiological studies for determining its size (a cone of angles 130°–135° vertical and 200°–220° horizontal) [Dag11]; in [CU90], it is demonstrated that there is a gradual dropoff in processing efficiency around the focus of attention: this pushed us in designing a VFOA with smoothed bounds (see the next section). Thanks to the results of [FUCH15; FUY15], we assume that the intersection of the VFOA with the scene indicates the probable future path, and, in the case of other people within the VFOA, they would be processed in determining possible colliding areas, which will be avoided with a certain probability.

It is worth noting that, experiments of the same kind of [FUCH15; FUY15] in the case of subjects forming groups are not traceable in the literature. This individuates an unexplored area of research for the sociological field, since people that walk together would probably have a strongly different fixation behavior with respect to single subjects; in facts, people in a moving group, other than the individual fixations needed for path planning, need to keep a reciprocal eye contact to maintain the social connection, that is, managing the turns in a conversation, processing non-verbal social signals etc. [Ken67; Ken90]. Because of this,

people in groups are considered in this paper as a single subject (they should share a very similar trajectory, with similar destination), with an extended VFOA obtained as the merge of their individual VFOAs.

2.2 Trajectory forecasting

Trajectory forecasting [BHHA18; MT08] has been traditionally addressed by approaches such as Kalman filter [Kal+60], linear [MN89] or Gaussian regression models [QR05; Ras06; WFH08; Wil98], auto-regressive models [Aka69] and time-series analysis [Pri81]. The main limitation of these approaches is the lack of modelling the human-human interactions [ABW06; CS12; CS14; LPR11; TCP06], that instead plays an important role. More recent approaches have proposed to use convolutional neural networks [HLZHW+16], generative models [GJFSA18b] and recurrent neural networks [AGRRF+16] for modelling the trajectory prediction, as well as, the human-human interaction. In addition, the head pose orientation [DR12; HSTDC+18] has been utilized for trajectory forecasting.

Below, we group the related work into four categories and discuss the related approaches.
Human-human interactions. Helbing and Molnar [HM95] have considered for the first time the effect of other pedestrians to the behavior of an individual. The pioneering idea has been further developed by [LCL07], [MHLK17] and [PESV09], who have respectively introduced a data-driven, a continuous, and a game theoretical model. Notably, these approaches successfully employed the essential cues for track prediction, such as the human-human interaction and people intended destination. More recent works encode the human-human interactions into a "social" descriptor [ARF14] or propose human attributes [YLW15] for the forecasting in crowds. More implicitly, related methods [AGRRF+16; VS17] embed the proxemic reasoning in the prediction by pooling hidden variables representing the probable location of a pedestrian in a LSTM. Our work mainly differentiates from [AGRRF+16; LCL07; PESV09; VS17] because we only consider for interactions those people who are within the cone of attention of the person, (as also verified by psychological studies [IC01]).
Destination-focused path forecast. Path forecasting has also been framed as an inverse optimal control (IOC) problem by Kitan *et. al.* [KZBH12]. The follow-up works [AN04; ZMBD08] have adopted inverse reinforcement learning and dynamic reward functions [LK16] to address the occurring changes in the environment. We describe these approaches as destination-focused, because they require the end-point of the person track to be known. To eliminate this constraint, similar works have relaxed to a set of plausible path ends [DRS11; MHB16]. Unlike, our approach does not require this information to function.
Head pose as social motivation. Our interest into the head pose stems from sociological studies such as [Cv80; DR12; FUCH15; FUY15; FWK11; PV03; VCDPL13], whereby the head pose has been shown to correlate to the person destination and pathway. Interestingly, the correlation is higher in the cases of poor visibility, such as at night time, and in general when the person is being busy with a secondary task (*e.g.*, bump avoidance) further to the basic walking [FUCH15; FUY15].

In our experimental studies, we observed that the head pose is correlated with the movement, especially at high velocities, while slowing down this correlation decreases too, but still remaining statistically significant. These studies motivate the use of the head pose as proxy to the track forecasting. Although the image resolution is small in our problem, there are many approaches that perform real-time head pose estimation [BO04; GMHC06; HTGDC17; LYO15b; RR06a; SFYW99; TSCM13b]. In our experiments, we evaluate different head pose estimation approaches.
LSTM models. LSTM models [HS97] have been employed in tasks where the output is conditioned on a varying number of inputs [GDGRW15; VTBE15], notably hand writing

generation [Gra13], tracking [CADNT17], action recognition [DWW15; LSXW16], future prediction [HLZHW+16; LCVCT+17; SMS15] and path prediction [XHR18].

As for trajectory forecasting, Alahi *et. al.* [AGRRF+16] model the pedestrians as LSTMs that share their hidden states through a "social" pooling layer, avoiding to forecast colliding trajectories. This idea has been successfully adopted by [VS17]. In [SAS17], it has been extended for modeling the tracking dynamics. A similar approach [SDZLZ16; SZDZ17] has been embedded directly in the LSTM memory unit as a regularization, which models the local spatio-temporal dependency between neighboring pedestrians. In this work, we propose a variant of the social pooling by considering a visibility attentional area, driven by the head pose.

In most of the cases, the training of LSTMs for forecasting minimizes the negative log-likelihood over Gaussians [AGRRF+16; VS17] or mixture of Guassians [Gra13]. In general, when it comes to Gaussian log-likelihood loss functions, only bidimensional data (i.e. (x, y) coordinates) have been considered so far, leading to the estimation of 2 x 2 covariance matrices. These can be optimized without considering the positive semidefinite requirement [Gra12], that is one of the most important problems for the covariances obtained by optimization [PB96]. Here, we study the problem of optimizing Gaussian parameters of higher dimensionality for the first time.

Chapter 3

Object Detection

3.1 Overview

Object detection is defined as task of localizing instances of real world objects such as cars, chair, person, desk etc. as shown in Fig. 3.1. Predominantly, the localization is performed by bounding boxes but in some areas it also accomplished by finding points close to the centers of the objects, or drawing ellipses. Majority of the research in object detection focuses on finding objects that occur more frequently than others in our daily lives such as (faces, pedestrian, cars, road sign etc.) due to the wide spread application of such objects.

However, generic object detection is often regarded as an ill-posed problem, as definition of object potentially could be very subjective based on task. In some cases it is a very tedious and expensive task to annotate all instances of the objects in a scene such as (e.g house, doors, windows, chairs etc). Therefore, in many applications the term object is well constrained and pre- defined in terms of its appearance, shape scale etc [FMFGL+96]. Although, as discussed the term "object" could be subjective therefore it has changed and evolved from one task to another and among different public datasets.

All recent approaches in object detection that perform well are learning based architectures. Which mean that all of these approaches would require huge amount of annotated data to be trained properly. In the last decade, a lot of attention has been given to the benchmarks in object detection. PASCAL VOC [EZWV06], ImageNet challenge [RDSKS+15] and MS COCO datasets [LMBHP+14] are currently the most popular and widely used benchmarks in general object detection. These benchmarks contain diverse set of object categories such as person, car, plants, animals, aeroplanes etc. More specifically PASCAL VOC has 20 object categories, ImageNet has 200 and MS COCO contains 91 object classes.

Object detection has a wide spread applications, including image retrieval, object counting, retrieval of items from warehouses, mail sorting, video surveillance, autonomous driving, robotics, detecting apparel and detecting logos of popular brands. Object detection also plays a vital role in other computer vision research areas, such as object tracking, object segmentation, caption generation and visual reasoning.

Due to wide spread application, different paradigms of object detection have been explored. Initially, in object detection the concepts from signal processing, such as auto corelation and template matching were used. Soon these concepts were taken over by 3D shape based CAD models(cite). Besides, computational overhead, the problem of texture and objects occurring in different scales were the bottleneck of these approaches. In the last couple of decades, object detection saw a tilt towards part based models[FGMR10]. These methods operated in a sliding window manner scanning through all of the image, naturally all the methods despite achieving relatively reasonable performances on several benchmarks were computationally expensive and did not always scale up.

In the recent times, the most popular paradigm in object detection is object proposals. An object proposal is a candidate for an object detection and/or segmentation. Object proposals enable the subsequent analysis stages to focus on a small set of image regions. They need

FIGURE 3.1: Example of object instance localization and recognition. Usually, an object detection framework, outputs bounding boxes along with object instance label and confidence score. Illustration take from [RHGS15b]

to have high recall, with a corresponding candidate for as many objects as possible, without increasing the total number of proposals too much. They have played a significant role in object detection methods during last decade by replacing sliding window approach and enabling the use of more advanced classifiers (Uijlings et al. 2013 [UVGS13]).

3.2 Faster R-CNN for General Object Detection

Given the success of image classification results obtained by deep networks combined with selective search a robust candidate generation method, [GDDM14] were the pioneers of R-CNN based object detection. The key idea was to combine region proposals with features obtained by CNNs for object detection. Initially, they adopted AlexNet [KSH12] along with Selective Search[UVGS13] as region proposals method and proposed a multi-stage training framework which outperformed all previous approaches by a significant margin. However, including proposals from an external source meant that it was not end to end trainable. Secondly, proposal generation was seen as an expensive operation and was regarded as the bottle neck. Motivated by [ZKLOT14; ZKLOT16], which illustrated how CNNs can be used for object localization, [RHGS15b] proposed Faster R-CNN, which had Region Proposal Network (RPN) along with discriminator Fast R-CNN[Gir15]. As illustrated in the figure 3.2, the main contribution of the work was a single network that generated proposals and had a Fast R-CNN region classifier. The convolutional layers, as illustrated in 3.2 were shared between Fast R-CNN and RPN (generating proposals almost free of cost), the whole framework was

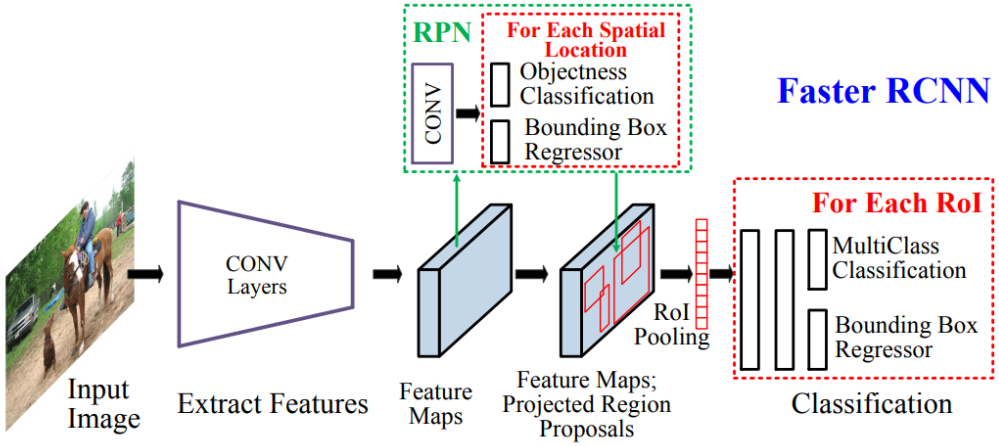


FIGURE 3.2: Faster R-CNN, complete diagram. Illustrations adapted from [LOWFC+18]

end to end trainable. Faster R-CNN defined new state of the art on all public benchmarks, to date it is considered as one of the popular paradigms of modern day object detection.

As shown in 3.2, RPN takes as input an image and outputs the classification score and bounding box coordinates (proposals). In order to generate proposals, a small network slides over the feature map of the last shared convolutional layer. As discussed in [RHGS15b], at each location, a spatial window (nxn) of the input feature map is fed into two fully connected layers, one for bounding box regression and the other for background vs. foreground classification score (objectness score). Importantly, at each location of the sliding window, RPN generates k proposals (4 k coordinates). These k proposals are paremeterized relative to k anchor boxes, as shown in 3.3. Authors in their experiments used 9 anchor boxes, 3 scales and 3 aspect ratios. Finally, class agnostic region proposals are then used by the discriminator Fast R-CNN which further refines proposals and assign each proposal a class category or label it as background. Faster R-CNN achieved top performances on PASCAL VOC, using 300 proposals per image and it takes 5 frames per second on a GPU for inference.

Original loss of Faster R-CNN as expressed in Eq. (3.1) is defined in the following equation. where L_{cls} and L_{loc} are the loss functions for background *vs* foreground classification and bounding-box regression respectively. In the next chapter, we will discuss how we modify this loss to enable person detection and head pose estimation. Initially, it had alternate training approach, however in recent times some amelioration [RHGS17] of the original frame work enabled it to be trained in a single step.

$$\mathbf{L}(p,u,t^u,v) = L_{cls}(p,u) + \lambda L_{loc}(t^u,v) \tag{3.1}$$

3.2.1 Faster RCNN for Person Detection

In computer vision, pedestrian detection is usually addressed as a separate problem than generic object detection [ZLLH16]. Despite, the success of deep learning based methods such as Faster R-CNN for general object detection, it seemed that they performed poorly for pedestrian detection. Zhang et.al 2016 [ZLLH16], were among the first one to investigate the reasons on why Faster R-CNN did not preform well when applied on pedestrian detection. Summarizing, they found at that RPN, actually is accurate in terms of recall, it is the discriminator Fast R-CNN, that degrades the performance. This was down to two main factors,

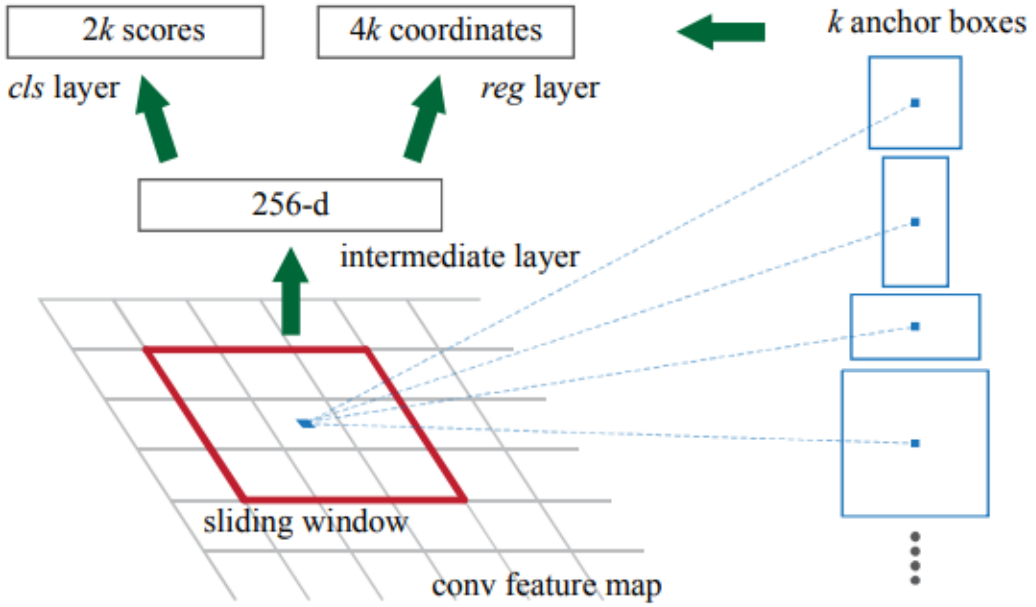


FIGURE 3.3: Region Proposals Network. Images taken from [RHGS15b]

1) insufficient resolution of the feature maps to handle small instances, 2) lack of hard negative mining and strong class imbalance between foreground and back ground. Furthermore, [ZLLH16] proposed an effective baseline that used a trous trick [CPKMY18] to increase the resolution of the feature map, along with a boosted forest [FHT+00; AFDP13] on top of RPN to effectively handle hard negative mining as shown in the 3.4.

Despite achieving decent performances, RPN+BF was missing a key component, it could not be optimized in a closed form since it was a cascaded framework. To overcome aforementioned issue, [ZBS17], proposed a pedestrian detector completely based on Faster R-CNN which with some minor modifications to the network and hyper parameters. Primarily, [ZBS17], modified four aspects of Faster RCNN. 1) they proposed better anchors by analyzing scales of the pedestrians on the training set. 2) Up sampling the input image to 2x. 3) By analyzing the average width and height, which was 40x80 respectively. It was evident that the default stride on VGG16 which was 16 pixels was too big for small scale pedestrian detection. Therefore, it was reduced to 8 pixels to handle small pedestrians. 4) They used adam solver[KA15] instead of SGD and 5) removed regions which were labelled "ignored" from training of RPN.

The modifications proposed by [ZBS17] to vanilla Faster R-CNN made it outperform all state-of-the art approaches on person detection by a considerable margin. In the upcoming chapter, using these modification along with a novel branch for head pose estimation along with person detection will be discussed in detail.

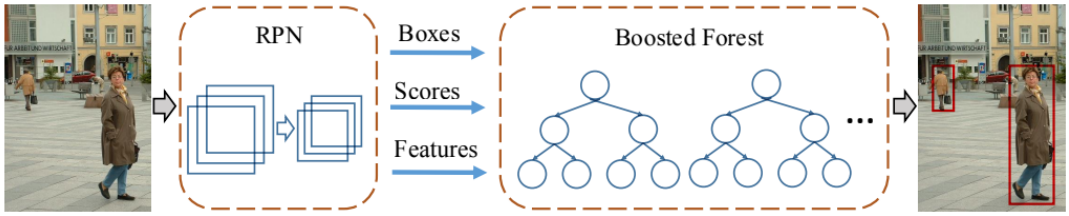


FIGURE 3.4: Complete pipeline of RPN+BF.Feature maps pooled from RPN are fed into cascaded boosted forest, for accurate pedestrian detection. Images taken from [ZLLH16]

Chapter 4

Head Pose Estimation

4.1 Introduction

The head pose is an important visual cue for several computer vision applications. In surveillance videos, the joint attention of people towards a direction can signal a particular event is happening [GXH10]. In social signal processing, the head orientation is necessary to infer group formations [CBPFT+11a] and capture social roles, such as leaders/followers [Eng94]. Most recently, the head pose has been used for novel marketing strategies and architectural design, as a proxy to personal interest in goods, impact of adverts and space utilization [DLB10].

The head pose estimation (HPE) problem is challenging in particular when people are captured at far and not yet addressed "in the wild". In many practical problems, such as video surveillance, HPE input is a head region as small as a 24×24 head pixel. This information alone is not enough to obtain reliable performance in HPE [TSCM13a], and multi-view camera setting are necessary [RSRVL+14].

This paper proposes to increase HPE performance by leveraging information from the entire body of the person instead of using the head information only.

Specifically, we enrich the recent Faster RCNN [RHGS15a] architecture with a branch specialized on the yaw modeling of the head pose (in this work, we focus on yaw, keeping the modeling of pitch and roll as future goals), called Head Pose Network (HPN). The idea is to jointly optimize the pedestrian detection and the HPE tasks, in order to establish and exploit a structural connection between the appearance of the body and the head pose. Secondly, we manually label the Town Center dataset [BR09b], which nicely portrays a surveillance scenario where 71,446 heads are imaged on 24×25 pixel patches.

The experiments, on this dataset and on standard benchmarks (oracle head detections are provided) show the net potentialities of our approach; additional ablation studies confirm that the body estimation, even if noisy, greatly improve the head pose estimation.

4.2 Datasets

The Town Center dataset [BR09b] has 4,500 frames portraying a crowded scenario with an average of 16 pedestrians per frame. The average size of the heads is about 24×25 pixels. We enrich the pedestrian bounding boxes labels by manually annotating the head direction. Towards this goal, we developed a software with a point-click interface that allows the annotator to inspect few frames of the dataset, selecting the direction where the pedestrian is looking at.

From the annotation, we extract quantized head pose directions, namely, 4 and 8. We then divide the sequence into a training and a testing sub-sequence of length 3,000 and 1,500 frames respectively.

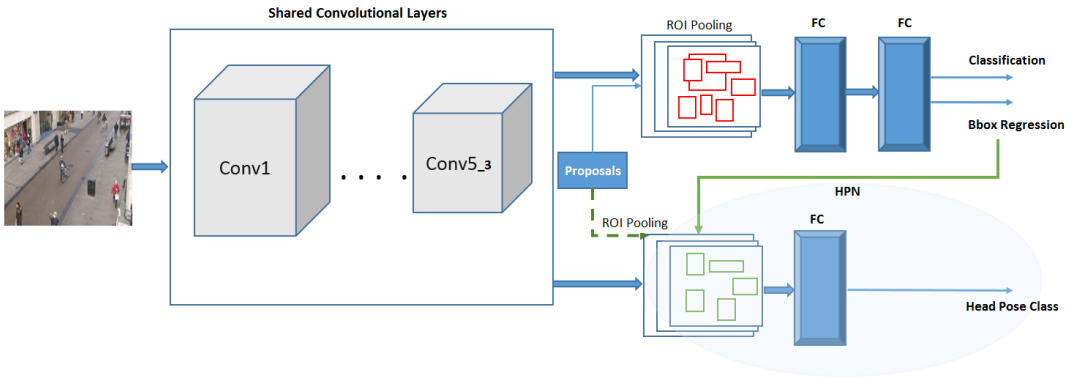


FIGURE 4.1: Network Architecture. The figure illustrates the proposed Head Pose Classification Network (HPN). The green dotted-line represents the filtered proposals at the training time and green solid represents the pedestrian detections at testing time.

QMUL [OGX09] and its extension with background class QMULB [OGX09]. *HIIT* dataset has 24,000 images with 6 head poses and a static background. *QMUL* dataset contains 15,660 images that has 4 different head poses with varying illumination and occlusion. *QMUL* dataset with additional 3,099 background images is referred to as *QMULB*.

4.3 Proposed Methodology

Our goal is to automatically predict the head pose of the pedestrians in addition to their bounding boxes. To this end, we propose a new network branch called the head pose classification network (HPN) as shown in Figure 4.1. The network is based on Faster RCNN [RHGS15a] but with novel additions and modifications to the network structure. Similarly to Fast RCNN, HPN has also two modules: a fully convolutional region proposal network (RPN) that provides class-agnostic object proposals and a Fast RCNN [Gir15] approach classifying the incoming proposals into pre-defined object classes.

In our HPN approach, we add an additional branch to the Faster RCNN network after the last shared convolutional layer (*i.e.* conv5_3), parallel to the classification and regression layers of the Faster RCNN. HPN includes also its own ROI pooling layer, a fully connected layer with sigmoid activation, and a K-way softmax layer for view-frustum classification for K discrete classes.

4.3.1 Training

We keep the alternative optimization approach as described in the Faster RCNN approach [RHGS15a] which iteratively trains the RPN and Fast RCNN stages. Related to the RPN optimization, we keep the shared convolutional layers of Faster RCNN in their original form. Moreover, the default Fast RCNN specific layers remain unchanged. The ROI pooling layer of the original Fast RCNN takes each object proposal as input and extracts a fixed-length feature vector from the entire feature map which is then fed into a couple of fully connected layers (fcs). Our new ROI pooling layer of HPN works in the same way, except it takes only filtered region proposals at the input. This is important since we want to learn the head pose of the pedestrian proposals without being distracted by the pedestrian false-positives. To select the examples for training the HPN, we use the standard Jaccard overlap of greater than or equal to 0.5 between the ground-truth bounding boxes and the region proposals.

Adding this parallel branch (HPN) in the Fast RCNN framework essentially extends the multi-task loss of Fast RCNN to penalize the view-frustum of the person bounding box. This allows us to learn jointly both detection and head pose classification tasks. Following the same naming conventions as Fast RCNN paper, our multi-task loss for jointly training pedestrian detection and head pose is given by,

$$\begin{aligned} L(p,u,t^u,v,h,g) =& L_{cls}(p,u) \\ &+ \lambda[u=1]L_{loc}(t_u,v) \\ &+ \gamma[u=1]L_{hp}(g,h) \end{aligned}$$

where L_{cls} and L_{loc} are the original loss functions for background *vs* pedestrian classification and bounding-box regression respectively. We refer the reader to original paper [Gir15] for more details on these terms. The L_{hp} term refers to the loss for the head pose of the pedestrian. We are using the softmax loss over K discrete directions of the head pose. Here g is the ground-truth label of the head pose class and $h = (h_1, h_2, ...h_K)$ is the output vector of softmax probabilities. Hence, $L_{vf} = -\log h_g$, is the negative log loss for the true view-frustum class g. As mentioned earlier, we train only for positive head pose classes and do not introduce any background class. This is given by the Iversion bracket indicator function $[u = 1]$. This means the two losses L_{loc} and L_{vf} are only used when the region proposals correspond to the pedestrian class. These losses are ignored for the background proposals. The weights λ and γ of the later two tasks are hyper-parameters which are set to 1.0 in our experiments.

4.3.2 Testing

At test time, our approach works in three stages. First the RPN outputs object proposals and passes them on to Fast RCNN detection network as usual. Note that this procedure basically is the Faster RCNN framework where we keep the pedestrian detections of the Faster RCNN with the confidence score 0.5 or greater. Finally, our HPN predicts the view-frustum class for each of these incoming detections.

4.3.3 Head pose classification

We first show the behavior of our technique in detecting and classifying head poses starting from raw frames. At the same time, we include ablation studies analyzing performance on head pose estimation. The latter test assumes that the head has been already detected by an oracle. The comparative approaches will be introduced later in the section.

Methods	**Dataset**	*HIIT*			*QMUL*			*QMULB*		
	Image Size	15x15	20x20	50x50	15x15	20x20	50x50	15x15	20x20	50x50
Frobenius [TSCM13a]		82.4	89.6	95.3	59.5	82.6	94.3	54.5	76.5	92
CBH [TSCM13a]		84.6	90.4	95.7	59.8	83.2	94.9	57	76.9	92.2
RPF [LYO15a]		97.6	97.6	97.6	94.1	94.3	94.3	91.9	92.1	92.2
PSMAT [OGX09]		-			-		82.3	-		64.2
ARCO [TFSMC10]		-			-		93.5	-		89
HPN		**98.4**	**98.9**	**99.01**	**97.4**	**97.9**	**98**	**95.3**	**95.9**	**94.7**

TABLE 4.1: Comparison of head pose classification accuracy in regard to image scale variation.

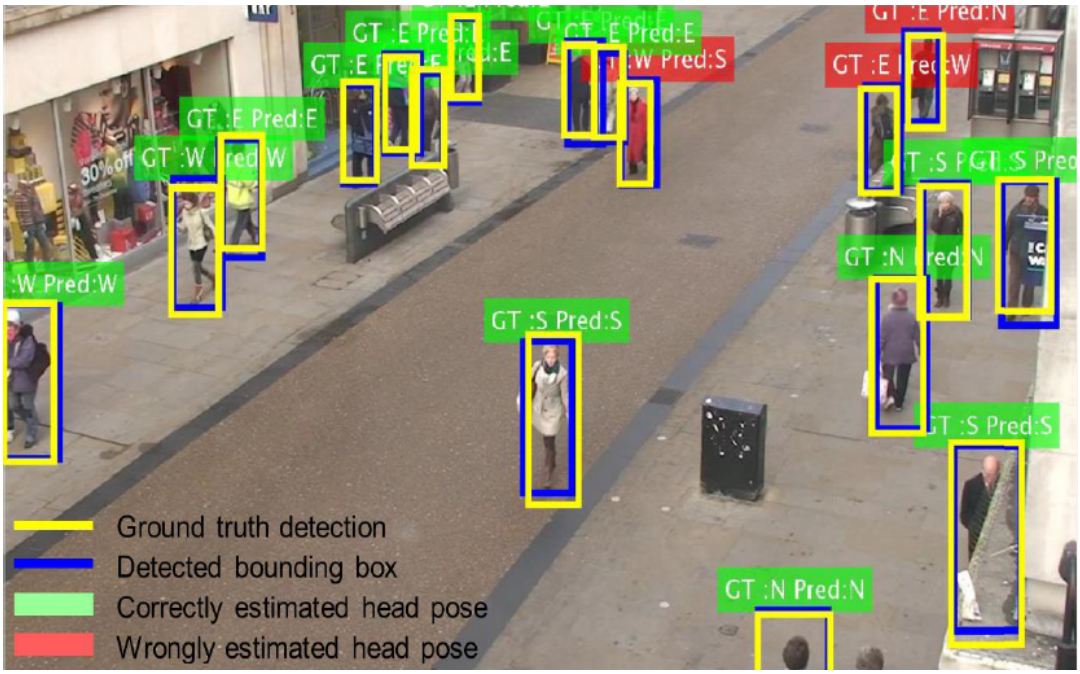


FIGURE 4.2: Qualitative results of our proposed model. Jointly detecting people and estimating their head pose.

4.3.4 Head pose estimation in the wild

The protocol for evaluating the pose estimation in the wild assumes that the algorithm takes a frame as input, and provides pedestrian bounding boxes plus the head orientation, initially evaluated over 4 classes (north, east, west, south) as shown in the Fig. 4.2. Additionally, we also pose head pose estimation as a regression problem as shown in Fig. 4.3. Results are in Table 4.2. We report LAMR (Log Average Miss Rate) [DWSP09] and AP (average precision) [EVWWZ10] for monitoring the pedestrian detection performance. It is worth noting that, in the head pose estimation accuracy, missed heads are counted as wrong detections: in this way, false negatives in the pedestrian detection flow down and impact in the final score. False positives are captured by LAMR and AP scores.

As competitors, we evaluate the Faster R-CNN [RHGS15a] directly as head pose estimator in the wild, trained over pedestrian bounding boxes associated to 5 classes (4 head directions and a background class, *FR-CNN 5-class* in the table 4.2). This will help us in showing the added value of our HPN branch in the joint optimization, which is absent here. The poor LAMR score (78%) contrasts the rather positive AP score 0.81%. The pose estimation accuracy, based on the whole body, achieves a reasonable 66%.

The second alternative approach is composed by a recent head detector, the Face detection with Aggregate Channel Features (FACF) [YYLL14a], which has shown to work pretty good on raw images, plus a head pose estimator, the Random Projected Forest (RPF) [LYO15a], which takes as input head bounding boxes, *FACF + RPF*. Both of them have been trained on the training partition of our dataset. As visible, performance is dramatically inferior, since obviously the head patches are very tiny and hard to catch without the body context.

The third approach wants to fill this gap, adding a pedestrian detection to constraint the head detector to work on pedestrian bounding boxes. In this case, we consider the Local Decorrelation Channel Features detector (LDCF) [NDH14], giving rise to the *LDCF + FACF + RPF* pipeline. Results on Table 4.2 show that performances are higher, but still inferior than FR-CNN 5-class.

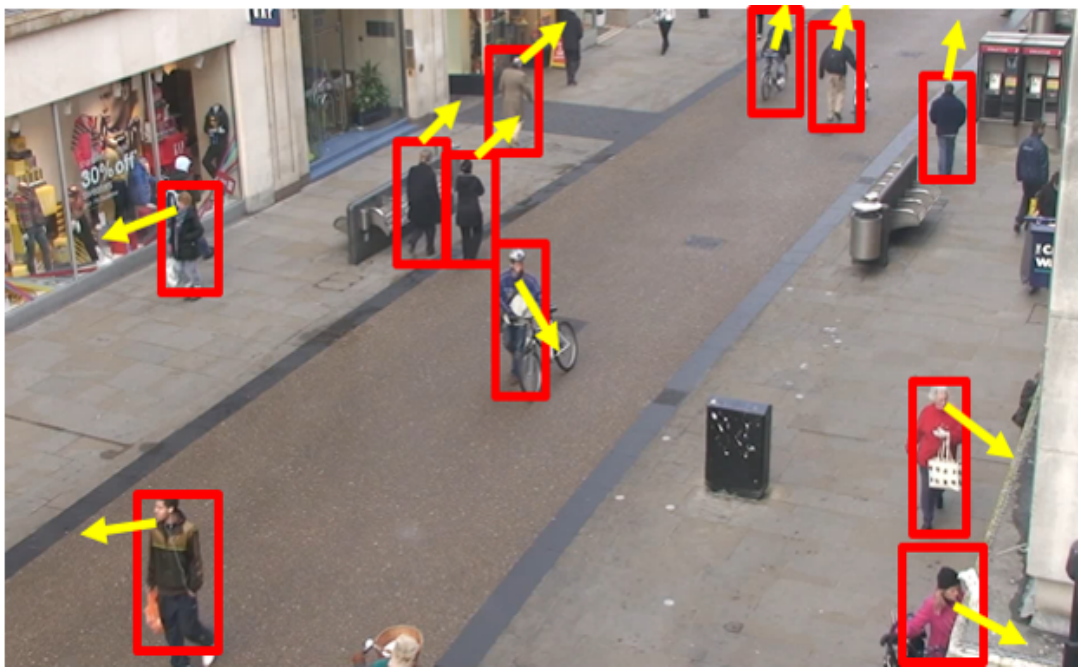


FIGURE 4.3: Regressing the head pose of the person in a real world surveillance scenario.

We further question the importance of face detection by testing *LDCF + HRCNN + RPF*, where a CNN-based head detector (HRCNN [JL16]) replaces the FACF. Reasonably, HRNN improves the head detection considerably, 10% LAMR and 12% AP (cf. Table 4.2), resulting in a better but still poor head pose estimation score of 50%. We conclude from this that the face, when so tiny, is not sufficient to estimate the pose estimation alone.

We mark as "ours" in the table the combination of pedestrian detection and pose estimation, jointly optimized within our model, cf. Eq. 1. As seen from Table 4.2, in the Town Center dataset a Faster-R-CNN person detector performs on par with the person specific LDCF [NDH14]. More interestingly, using the whole body for the estimation of pose greatly improves performance by 18%, resulting in the best technique, HPN, which we propose. This resonates with the baseline Faster-R-CNN 5-classes in the first row, also based on the whole body.

We compare performances for pedestrian detection of our joint framework against FR-CNN N-Cls baseline, and state-of-the art pedestrian detector LDCF [NDH14], FCF [ZBS15] and Faster RCNN [RHGS15a]. We report our results in Table 4.4 . In Our framework, we investigate the VGG model . We re-train VGG model, which was originally pre-trained on ImageNet, for baseline, Faster R-CNN and our joint framework. We use the standardize matrices for pedestrian and object detection (LAMR, AP). Faster RCNN, performs best for pedestrian detection closely followed by ours joint model. However, when Faster RCNN tries to incorporate the information about head pose(FR-CNN-N-CLS) the performance drops significantly. Stressing the fact that ours joint model, incorporates the simultaneous people detection and head pose estimation in a more reliable fashion.

4.3.5 Ablation study: head pose classification

The ablation studies serve to evaluate how our approach works in the case of a correct person detection. To enrich the analysis, in Table 4.3 we consider different numbers of pose quantization, namely 4 and 8 classes, in which the quantization has been obtained by uniformly

dividing 360 degrees. As competitors, we consider *RPF* [LYO15a], the FR-CNN N-class (N refers to the quantization bins), and 2 different versions of our approach. The variation we want to analyze (***Ours disjoint optimization***) does the following thing: as in the proposed version, the complete body is used for head pose estimation but the optimization terms for object classification L_{cls} and bounding box regression L_{loc} are set to zero. In practice, this breaks up the joint optimization and let the system operate as two separate modules, where the object detection loss does not contribute to the head pose classification training.

Ours Joint Model is the proposed methodology where as explained above pedestrian detection module and the head pose estimator are jointly optimized as the output is shown in 4.2. Table 4.3 illustrates the robustness of our approach in regards to the granularity level of head pose. Secondly, pedestrian detection and head pose estimation are related task, therefore when posed as a joint optimization problem performance for head pose estimation gets boosted.

In Table 4.1 the second ablation study stresses the ability of our approach in estimating the head poses by starting from correct head bounding boxes. For this purpose we train HPN over head images and pose it as a classification problem. Except for the QMULB [OGX09] dataset, which has an additional background class, in that case we train HPN to have a cascaded output, first distinguish between person and non-person and then classifying only persons for the head poses. This procedure is consistent to our proposed joint model. Results have been computed on the datasets HIIT [TSCM13a], QMUL [OGX09] and its extension with background class QMULB [OGX09]. *HIIT* dataset has 24,000 images with 6 head poses and a static background. *QMUL* dataset contains 15,660 images that has 4 different head poses with varying illumination and occlusion. *QMUL* dataset with additional 3,099 background images is referred to as *QMULB*.

Finally, as shown in Table 4.1, proposed HPN is capable of overcoming in terms of average accuracy, all of the competitors at each resolution.

Pipeline	**Pedestrian Detection**		**Head Detection**		**Head Pose Est. Accuracy**
	LAMR	AP	LAMR	AP	
FACF [YYLL14b] + RPF [LYO15a]	N/A	N/A	90.67	0.336	0.3
LDCF [NDH14] + FACF [YYLL14b] + RPF [LYO15a]	54.99	0.83	96.37	0.2087	0.27
LDCF [NDH14] + HRCNN [JL16] + RPF [LYO15a]	54.99	0.83	84.36	0.31	0.5
Ours	**55**	**0.86**	N/A	N/A	**0.68**

TABLE 4.2: Head pose estimation in the wild. For LAMR, lower is better.

Method	**Classification Accuracy (4 classes)**	**Classification Accuracy (8 classes)**
RPF [LYO15a]	0.6	0.31
FR-CNN N-class [RHGS15a]	0.71	0.32
Ours (Disjoint Optimization)	0.72	-
Ours (Joint Model)	0.74	0.33

TABLE 4.3: Head pose classification accuracy on oracle.

Method	AP	LAMR
LDCF [NDH14]	0.83	54.9
FR-CNN N-Cls	0.81	78
Faster RCNN [RHGS15a]	0.87	52.3
FCF [ZBS15]	0.82	61.1
Ours	0.86	55

TABLE 4.4: Pedestrian detection results. For LAMR lower is better

Chapter 5

LSTM Overview

Recurrent Neural Networks has the ability to learn contextual information when mapping input and output sequences [Kaw08]. However, like any deep feed forward neural network RNNs are trained by propagating the error through time. Back propagating error through time makes the gradient either to decay or blow up exponentially as moves around the recurrent connections. These phenomenons are know as vanishing and exploding gradients problems [HBFS+01]. This problem can also be seen in the fig. 5.1

Ever since the discovery of the vanishing gradient problem, several remedies have been proposed such as initialize the network weights so that vanishing gradient is not pronounced or having Echo State Networks and Long Short Term Memory. In this thesis we will discuss about Long Short Term Memory networks [HS97].

5.1 Brief History of LSTM

Detailed evolution of LSTMs could be found [GSKSS17]. This chapter will briefly present major architectural changes that happened to LSTMs over the period of time. Earliest version of the LSTMs [HS97] included cells, input and output gates [GSKSS17]. However, these LSTM were missing peephole connections, unit biases or input activation. As also discussed by [GSKSS17], training was only done using the combination of Real Time Recurrent Learning (RTRL) [RF87; Wil89] and Backpropagation Through Time. Initially, only the gradient of the cell was backpropagated and all other gradients were truncated, naturally limiting the capabilities of the LSTM. However these practices were soon dropped in the favor of modern day LSTMs.

Forget Gate Soon after the emergence of LSTMs, it was soon realized that not all past information was useful to keep in LSTM. Therefore, [**gers1999learning**] proposed a mechanism for LSTMs, where they forget and reset to their own state. This gate is known as forget gate. In continual task such as trajectory forecasting, this aspect of the LSTM is vital.

Peephole Connections

In the case of time series data, precise timings are of paramount importance, original LSTMs could not perform it with accuracy as their was no mechanism that allowed cells to control the gates. [**gers2000recurrent**], were the first one to propose peephole connection where cells were connected with gates as shown in Fig. 5.2 (blue dots).

Full Gradient

Graves et al. 2005 [GS05], proposed the final modification to the modern day LSTMs. They were the first one to present full backpropogation through time. This full backpropoagtion through time allowed LSTMs to be more reliable and robust.

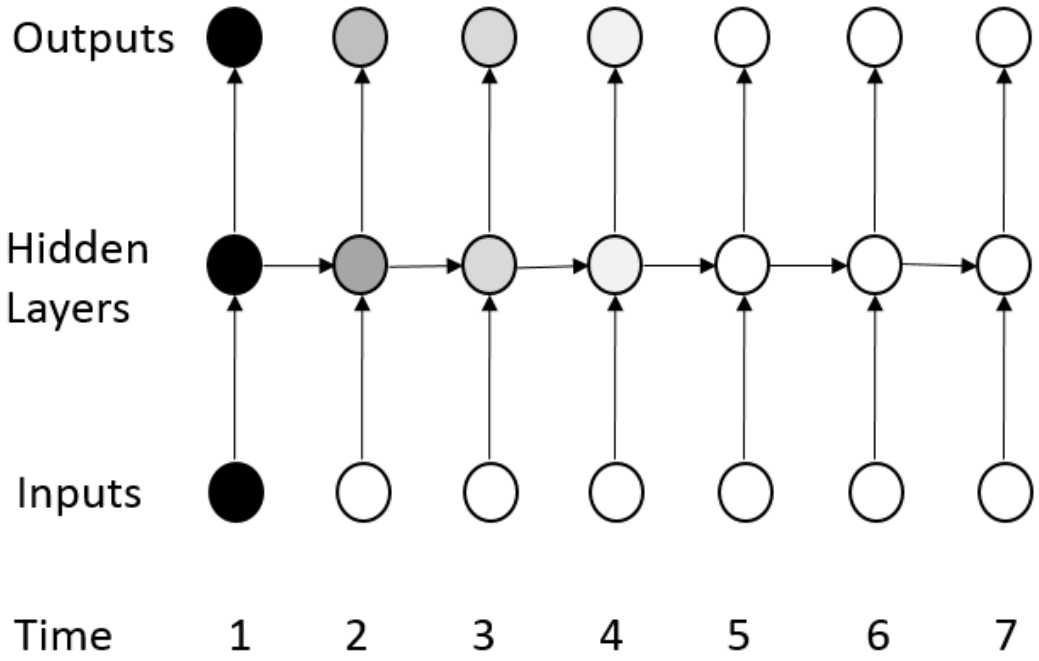


FIGURE 5.1: Illustration of the vanishing gradient problem in RNNs. One could see the effect of gradient vanishes over time (lighter color)

5.2 Social LSTM

Social LSTM [AGRRF+16], was among the pioneer works in the field of trajectory forecasting. It was a complete data driven approach, which used temporal capabilities of RNNs for pedestrian path prediction. Similar to previous works [Gra13] Social LSTM developed an LSTM based model for trajectory forecasting. On 5 public benchmarks Social LSTM outperformed all previous handcrafted energy based approaches by a significant margin.

In Social LSTM, one LSTM per person was proposed, as shown in Fig.5.4 LSTM model learns the representation of how an individual navigate through crowded spaces. However, we humans when we navigate we experience several forces such attractive (groups) and repulsive (collision). A naive LSTM model could not capture such human-human interaction. Therefore, a "communication" strategy was proposed by Alahi et al.2016 [AGRRF+16], named as social pooling.

Social pooling, as illustrated in Fig.5.4, pools the hidden layers of all neighbouring LSTMs (pedestrians) in the scene. This pooling operation basically communicates with the given subject that where are his neighbours and where they will be in future. This pooling mechanism is similar to how humans interact with the environment (avoiding collisions and engaging into human human interaction). The social pooling is based on the spatial location of other pedestrians.

As described in detail [AGRRF+16], the hidden state of the LSTM $\mathbf{h_t^i}$ learns the representation for the i th person in that particular scene. Subsequently, these hidden states are shared between neighbours and for a given subject a hidden tensor $\mathbf{H_t^i}$ is constructed. For a hidden-state having dimension D and a neighborhood size $\mathbf{N_o}$, hidden tensor $\mathbf{H_t^i}$ for $\mathbf{i^{th}}$ trajectory of size $\mathbf{N_o}$ x $\mathbf{N_o}$ x D is defined as in eq (5.1):

$$\mathbf{H}_t^{(i)}(m,n,:) = \sum_{j \in \mathrm{N}_i} \mathbf{1_{mn}}[x_t^j - x_t^i, y_t^j - y_t^i] h_{t-1}^j, \tag{5.1}$$

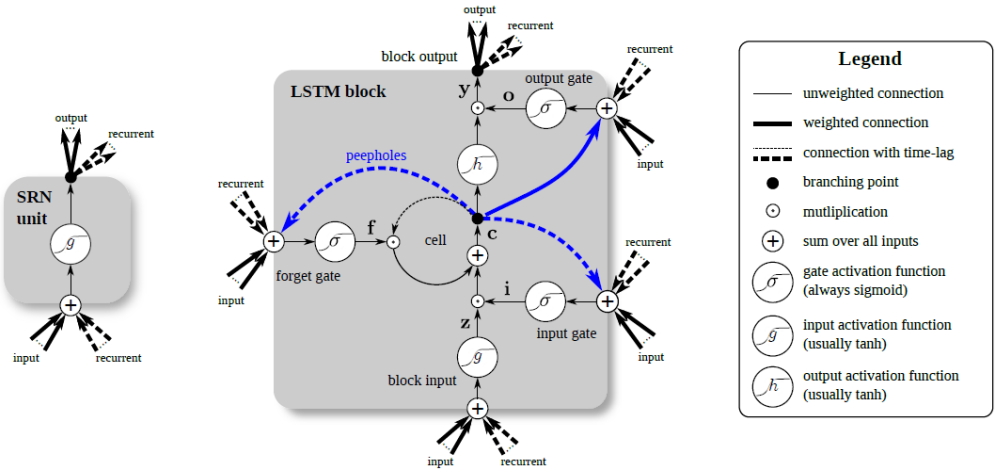


FIGURE 5.2: Detailed diagram of a Simple Recurrent Network unit and LSTM. Image courtsey [GSKSS17]

where $\mathbf{h_t^i}$ is the hidden state of the LSTM referring to the $\mathbf{j}$ th person at $\mathbf{t-1}$. $\mathbf{1_{mn}[x,y]}$ is an indicator function, which determines if (x, y) is in the (m, n) cell of the grid, and $\mathbf{N_i}$ are other individuals in the scene for teh subject $\mathbf{i}$. This pooling operation could be seen as in Fig. 5.4.

Each LSTM model is instantiated using equation (5.2). where the embedding function ϕ is the linear projection, via the embedding weights $\mathbf{W}$, into a D-dimensional vector, with D the dimension of the hidden space. This is followed by a ReLU activation function. Same transformations are applied to the embedding function of the hidden state equation (5.3)

$$\mathbf{e}_t^{(X,i)} = \phi(\mathbf{X}_t^{(i)}, \mathbf{W}_x). \tag{5.2}$$

$$\mathbf{e}_t^{(H,i)} = \phi(\mathbf{H}_t^{(i)}, \mathbf{W}_H). \tag{5.3}$$

Similar to graves et al. [Gra13], this work assumes a bivariate Gaussian distribution parameterized by μ, Σ $\mathbf{P}$

$$[\mu_t^{(i)}, \hat{\mathbf{\Sigma}}_t^{(i)}, \mathbf{P^i}_t] = \mathbf{W}_o \mathbf{h}_{t-1}^{(i)},$$

Finally, at training time, weights are estimated by minimizing the bivariate Gaussian log-likelihood for the each trajectory. The loss function is

$$L^i(\mathbf{W}_x, , \mathbf{W}_H, \mathbf{W}_o) \quad = \\ -\sum_{T_{obs}+1}^{T_{pred}} log\left(P(\mathbf{X}_t^{(i)}, \mu_t^{(i)}, \mathbf{\Sigma}_t^{(i)}), \mathbf{P^i}_t\right),$$

where $\mathbf{T_{obs}}$ is the last frame of the observation period, while $\mathbf{T_{obs}}$ + 1, . . . , $\mathbf{T_{pred}}$ are the time frames for which we provide a prediction. The loss of Eq. (5.4) is minimized over all the training sequences. To prevent overfitting, we additionally include an l2 regularization term.

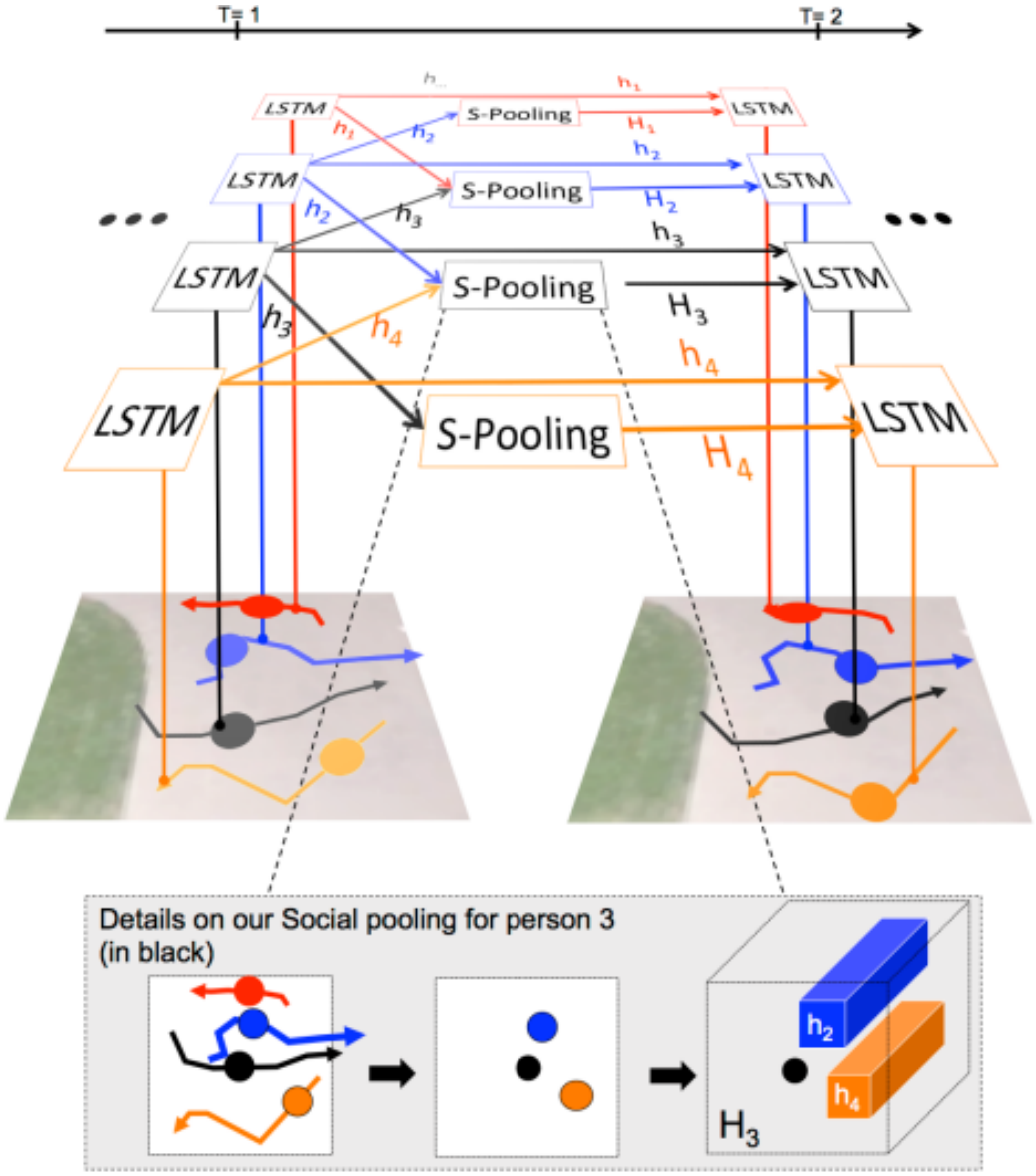


FIGURE 5.3: Description of Social LSTM. Authors proposed one LSTM per person. Image adapted from [AGRRF+16]

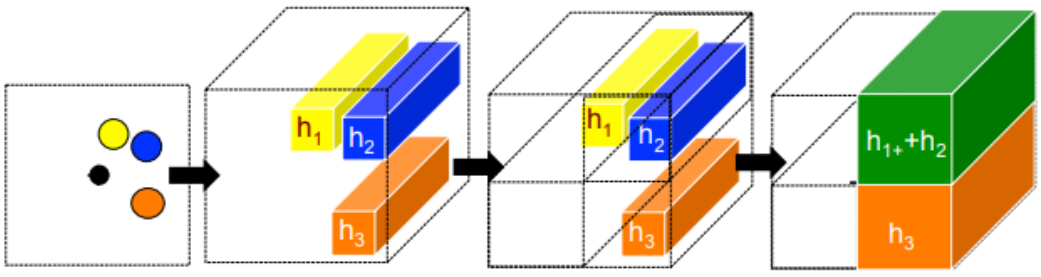


FIGURE 5.4: Description of Social LSTM. Authors proposed one LSTM per person. Image adapted from [AGRRF+16]

Chapter 6

Trajectory Forecasting

6.1 Related works

A large body of literature have addressed the topic of path prediction, by adopting Kalman filters [Kal+60], linear regressions [MN89], Gaussian regression models [QR05; Ras06; WFH08; Wil98], autoregressive models [Aka69] and time-series analysis [Pri81]. Our approach departs from these classical approaches because we also consider the human-human interactions and the person intention, expressed by the VFOA.

Human-human interactions. The consideration of other pedestrians in the scene and their innate avoidance of collision was first pioneered by [HM95]. The initial seed was further developed by [LCL07] and [PESV09], which respectively introduced a data-driven and a continuous model. Notably, these approaches remain top performers on modern datasets, as they successfully employ essential cues for track prediction such the human-human interaction and the people intended destination. More recent works encode the human-human interactions into a "social" descriptor [AGRRF+16; ARF14; MWF16] or proposes human attributes [YLW15; MWFF17] for the forecasting in crowds. Our work mainly differentiates from [LCL07; PESV09] because we only consider for interactions those people who are within the cone of interest of the person, which we encode with the VFOA (as also maintained by psychological studies [IC01]).

Destination-focused path forecast. Starting from the seminal work of Kitani *et al*. [KZBH12], path forecast has been cast as an inverse optimal control (IOC) problem. Follow-up work has additionally utilized inverse reinforcement learning [AN04; ZMBD08] and dynamic reward functions [LK16] to address the occurring changes in the environment. We describe these approaches as destination-focused because they all require the end-point of the person track to be known, which later work has relaxed to a set of plausible path end points [DRS11; MHB16]. We share with these works the importance of the person intention, but we believe that knowing the destination undermines the reason why we may be predicting the trajectories. By contrast, we represent the person intention by their VFOA which, as we show, may be estimated at the current frame.

VFOA and the social motivation. The interest into the VFOA stems from sociological studies such as [Cv80; DR12; FUCH15; FUY15; FWK11; PV03; VCDPL13], whereby VFOA has been shown to correlate to the person destination, pathway and speed. Interestingly, the correlation is higher in the cases of poor visibility, such as at night time, and in general when the person is being busy with a secondary task (*e.g*. bump avoidance) further to the basic walking. These studies motivate the use of VFOA as a proxy to forecasting trajectories. Using VFOA comes with the further advantage that it can be estimated [BO04; RR06a; SFYW99] on a frame basis, thus requiring no oracle information and enabling a real-time system. While our experiment is agnostic about the head pose estimation algorithm, in our experiments we will use an off-the-shelf head pose estimator [HTGDC17].

Sequences	Number of frames	Number of pedestrians	Pedestrian per frame	Avg traj.
UCY	5,405	434	32	404
Zara01	8,670	148	6	339
Zara02	10,513	204	9	467
TownCentre	4,500	230	16	310

TABLE 6.1: Dataset Statistics

6.2 Datasets

We evaluated our approach on publicly available benchmarks, UCY [LCL07] and TownCentre [BR11] and compared it against state of the art methods. The benchmark UCY contains three sequences showing two different scenarios. Zara01 and Zara02 sequences show a public street with shops and cars, the number of pedestrians is quite limited and the trajectories are somehow constrained since entry and exit points are in a limited portion of the image border. UCY sequence is taken in a university campus plaza and it shows a dense crowd moving in several directions without any physical constraint. Similarly, TownCentre dataset portrays a crowded real world city centre scenario. The four datasets have in total of 29,088 frames with 1,016 pedestrians. More details about each sequence are given in Table 6.1.

6.3 Proposed Approaches

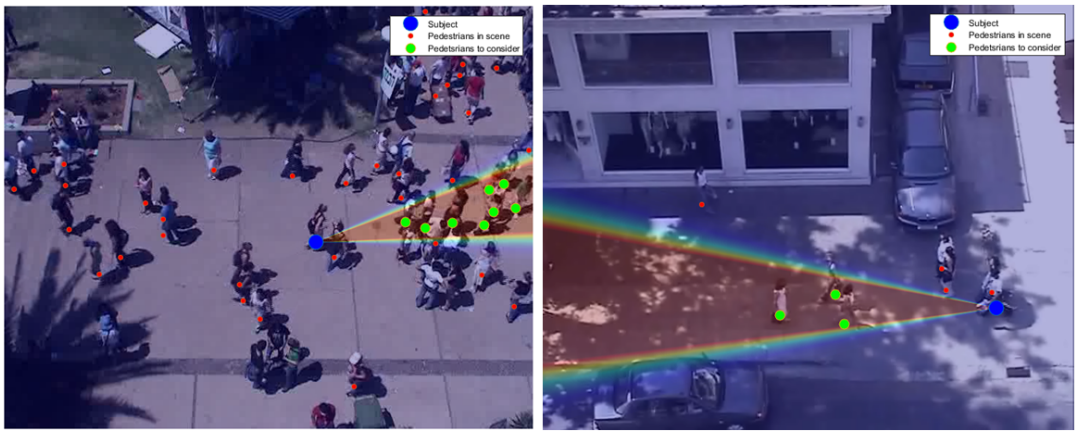


FIGURE 6.1: Graphical explanation on the selection of pedestrians to be taken into account for the avoidance term. The large blue dot represents the target pedestrian, the green dots are the pedestrians he/she tries not to collide to, and the small red dots are the pedestrians he/she is not aware of because out of the view frustum. (Best viewed in colors.)

We formulate the predictive model as a joint optimization problem, where the position of each individual in the next frame is simultaneously estimated by minimizing an energy function.Our loss function consists of three intuitive potentials: (1) a collision avoidance term, which accounts for the multi-agent nature of the system, (2) a destination term, which accounts for the goal of each individual behaviour, and (3) a constant velocity term. The general idea behind our model is that, when in an open space, a person walks towards a destination point trying to avoid collisions with other pedestrians and static objects. While doing this, she/he prefers to move smoothly, *i.e.* limiting accelerations both in terms of intensity and direction.

Our cost function has the general form:

$$\mathcal{C} = w_A \cdot E_A + w_V \cdot E_V + w_D \cdot E_D$$

where w_A, w_V, and w_D are weighting factors, and E_A, E_V, and E_D are the respective three energy terms discussed in the following.

Let us consider a video sequence of T image frames as $\mathcal{S} = \{I_t\}_{t=1\dots T}$. At each frame t, a set of N pedestrians are detected and their position on the ground plane is $P_i(t)$, $i = 1 \dots N$. For each individual, we define his/her head orientation $\theta_i(t)$. Finally, let us indicate with $\hat{P}_i(t+1)$ the predicted location of the individual i at frame $t+1$.

In order to promote smooth trajectories, we define the *velocity term* (E_V) as the summation over all the individuals' of the squared ℓ^2-norm of the acceleration vector:

$$\begin{aligned} E_V &= \sum_{i=1}^{N} \left\| \frac{d^2 P_i(t)}{dt^2} \right\|^2 = \\ &= \sum_{i=1}^{N} \left\| \hat{P}_i(t+1) + P_i(t-1) - 2P_i(t) \right\|^2 \end{aligned}$$

As for the *destination term* (E_D), we consider that a person is consistently looking at his/her short-term destination point while walking. Thus, this term is the additive inverse of the cosine of the angle comprised between the gaze direction θ_i, *i.e.* the head pose, and the direction of the predicted velocity:

$$E_D = -\sum_{i=1}^{N} \cos\left(\theta_i(t) - \underline{/\hat{P}_i(t+1) - P_i(t)}\right)$$

where $\underline{/\mathbf{v}}$ is the phasor angle of vector $\mathbf{v}$.

For the *avoidance term* (E_A), many different models have been proposed in the literature, mostly based on the concept of *social force* [HM95; PESV09; RSAS16; YBOB11]. The idea is that a person would not allow another individual to enter his/her personal space; thus, when walking, people adjust their velocity in order to avoid this kind of situations to happen. In this work we model the avoidance potential as a repulsion force that is exponential with respect with the distance between two predicted locations. Unlike many previous works, which consider the repulsion force only when 2 pedestrians are going to be closer than an isotropic comfort area, our method is more biologically motivated, assuming that the pedestrian reacts to what he senses in terms of sight, which is modeled by the VFOA. More formally, this term assumes the summation over all the individuals of the exponential of the minimum distance between the predicted location of the individual itself and the closest predicted location of another individual.

$$E_A = \sum_{i=1}^{N} e^{-\arg\min_j \, d_{ij}^*(t+1)} \quad , \text{ with } j \in \mathcal{F}_i(t), j \neq i$$

where $d_{ij}^* = \|\hat{P}_i(t+1) - \hat{P}_j(t+1)\|^2$, and $\mathcal{F}_i(t)$ is the set of all the individuals inside the VFOA of person i at time t. While in theory the view frustum is related to the gaze, we assume that in first approximation, in the scenario we are facing, the gaze is equal to the head orientation. Thus, we model the VFOA as a circular sector of angle 30°, where this last angle has been found experimentally (see in Sec. 6.5.7): surprisingly, this angle corresponds to the angle of the human focal attention [IC01], which can be likened to a "spotlight" in the visual receptive field that triggers higher cognitive processes like object recognition. A graphical explanation of the VFOA is given in Fig. 6.1.

Thus, the cost function of Equation 6.3 can be minimized with reference to $\hat{P}_i(t+1)$, $\forall i = 1 \dots N$. So at each step we predict the positions of all the pedestrians in the scene jointly. This optimization problem can be addressed with a *direct search method* for n-dimensional unconstrained spaces. The Nelder-Mead simplex method [LRWW98], adopted in this work, uses an iterative approach that maintain at each step a non-degenerate simplex of $n+1$ vertices, and updates the simplex according to the function value in the vertices. The method has a very low complexity, since it does not require to compute the gradient (as all the direct search methods) and typically requires the function evaluation on only one or two sample points at each iteration step.

6.4 Experiments

We evaluated our approach on publicly available benchmarks, UCY [LCL07] and TownCentre [BR11] and compared it against state of the art methods.

The evaluation protocol follows the most recent literature. We first downsample the frame rate of the videos of a factor of 10, resulting in a frame rate of 2.5 fps. Then, for each pedestrian detected, we predict their trajectory for the next 12 frames (4.8 seconds) by considering at every time step the predicted location of the target pedestrian and the ground truth positions of all the others. As for the evaluation metrics, we use the standard *Mean Average Displacement* (MAD) and the *Final Average Displacement* (FAD) error. The MAD metric is given by the average over all the pedestrians and all the frames of the Euclidean distance between the predicted location and the ground truth position. The FAD error is given by the average displacement of the 12-th predicted frame over all the trajectories.

6.4.1 Quantitative results

We compare our method with four state-of-the-art model-based approaches, namely Linear Trajectory Avoidance (LTA) [PESV09], Social Force model (SF) [YBOB11], Iterative Gaussian Process (IGP) [TK10], and multi-class Social Force model (SF-mc) [RSAS16]. We also provide results with a baseline method (Lin.) that merely estimates the next locations by using the previous velocity. For a fair evaluation, we need to point out that all the methods use different ground truth data and/or a priori information. All the approaches require the knowledge of the ground truth pedestrian position at each time step. In addition, IGP requires the exact destination point of each pedestrian (*i.e.* the last point of each trajectory, or the point where the pedestrian exits from the scene); LTA, SF and SF-mc require a soft version of the destination point, indeed they only need the direction the individual is pointing (*e.g.* North, South, East or West); SF and SF-mc also require to know which individuals are forming groups.

Differently, our approach does not require the knowledge of destination points or a direction but just the pedestrian position (as the others) and the labelled head orientation of each individual, no group membership is required. The destination point of each pedestrian, as well as other terms in the cost function Eq. 6.3 are then automatically estimated. We report sample model parameter in Table 6.4. Since head pose is crucial for forecasting, although people maintain a trajectory to their final destination, there might be the need to take short term deviations in order to avoid collision, obstacles or to engage in human-human interactions (*e.g.* a subject might take few steps in the complete opposite direction of the given destination point). This short term divergence is not addressed in any of the other methods and the head pose seems to be an effective mean towards this end.

Table 6.2 and Table 6.3 show that our method outperforms the state of the art methods in MAD, while it scores worst against the SF-mc on FAD in the UCY sequence. Please note that the comparison with IGP method with the FAD metric is not fair by definition, since it

Dataset	Lin.	LTA	SF	IGP	SF-mc	Ours
UCY	0.57	0.51	0.48	0.61	0.45	0.38
Zara01	0.47	0.37	0.40	0.39	0.35	0.30
Zara02	0.45	0.40	0.40	0.41	0.39	0.26
Town Centre	1.3	1.8	2.1	–	–	1.2

TABLE 6.2: Mean Average Displacement (MAD) error for all the methods on all the datasets.

Dataset	Lin.	LTA	SF	IGP	SF-mc	Ours
UCY	1.14	0.95	0.78	1.82	0.76	0.78
Zara01	0.89	0.66	0.60	0.39	0.60	0.59
Zara02	0.91	0.72	0.68	0.42	0.67	0.60
Town Centre	2.7	3.67	3.8	–	–	2.28

TABLE 6.3: Final Average Displacement (FAD) after 12 frames (4.8 seconds) for all the methods on all the datasets.

wA	wV	wD
0.1	1.16	1.0184

TABLE 6.4: Model parameters obtained from training sequences

Dataset	Ours (no frustum)	Ours
UCY	0.41	0.38
Zara01	0.31	0.30
Zara02	0.29	0.26

TABLE 6.5: Mean Average Displacement (MAD) with and without the view frustum condition in the avoidance term.

requires the annotation of the final point of each trajectory. Even with the unfair advantage for IGP, in a more densely crowded scenario like UCY, IGP performs poorly, since the short term divergence of a subject is much more prominent and is not addressed by the fixed destination point.

6.4.2 Ablation studies

It is worth noting that all approaches assume that a subject takes the next step accounting for all other pedestrians in the scene. This assumption is far to be true since in normal situations most people are unaware of what is happening behind themselves, and this does not effect their future movements. Thus, to prove the effectiveness of the the view frustum information, we conducted two ablation studies.

First, we turned off the frustum in the avoidance term, taking into account all the pedestrians in the scene. In such a case performances decrease of 2% in MAD and 5% in FAD, showing that the view frustum is beneficial for both metrics in all the sequences. (Table 6.5 and Table 6.6)

As a second experiment, we provided to the state-of-the-art approaches the destination points estimated frame-by-frame from the head pose. Results of Table 6.7, compared with the ones reported in Table 6.2, demonstrate how the use of head pose is beneficial also for other approaches, improving performances of LTA and SF of 5% and 6% on average respectively.

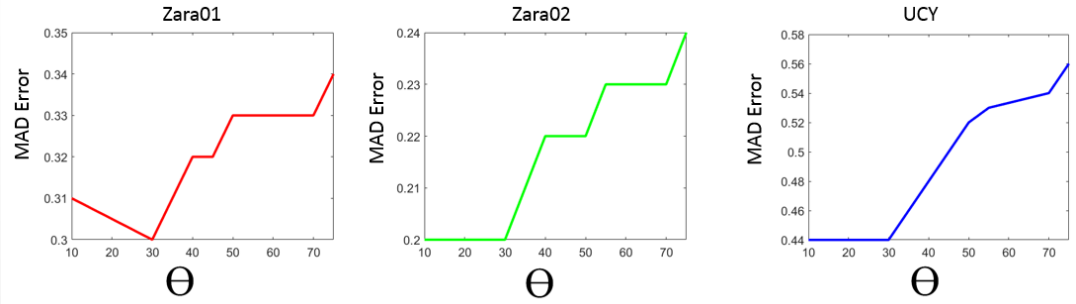


FIGURE 6.2: Θ angle of the VFOA in relation with the Mean Average Displacement error

Dataset	Ours (no frustum)	Ours
UCY	0.83	0.78
Zara01	0.65	0.59
Zara02	0.64	0.60

TABLE 6.6: Final Average Displacement (FAD) with and without the view frustum condition in the avoidance term.

Dataset	LTA	SF	Our
UCY	0.44	0.42	0.38
Zara01	0.33	0.32	0.30
Zara02	0.35	0.35	0.26
Town Centre	1.2	1.4	1.2

TABLE 6.7: Mean Average Displacement (MAD) for state of the art methods with destination point estimated from the head orientation.

Fig. 6.2 shows the study on the span of the Θ angle of the VFOA in relation with the MAD error, when the ground-truth head orientation is known. For this sake, we randomly sample 25 pedestrians per dataset (Zara01, Zara02 and UCY) and we compute the error while modulating Θ from 10 to 75 degrees with a step of 5. As visible in the figure, the range from 10 to 30 gives the best score, with 30 being the best absolute value. Actually, this does correspond to the angle defining the focal attention area [IC01].

6.4.3 Experiments with HPE

Once we have shown the theoretical advantages of our approach, we replace the oracle head orientation with the one estimated from a real head pose estimator [HTGDC17]. As most of the head pose estimators, the one used in this work outputs the head pose in a quantized format: dividing the 360° into 4 or 8 classes, thereby we also quantized the ground truth into the same format in order to understand the theoretical bounds that one could reach with the detector.

Looking at the results in table 6.8 we illustrate that even with the real head pose estimator, we could get competitive results with all the state-of-the-art approaches, which relies on strong ground truth information, highlighting the pragmatism of our approach. Additionally, by quantizating the ground truth we further illustrates that given an accurate pose estimator one could outperform the current state-of-the-art approaches. Moreover, as it can be noticed, finer granularity for head pose estimation proves to be more suitable in trajectory forecasting.

Dataset	GT	GT(4)	GT(8)	HPE(4)	HPE(8)
UCY	0.38	0.44	0.43	0.52	0.50
Zara01	0.30	0.39	0.37	0.44	0.42
Zara02	0.26	0.35	0.34	0.39	0.38
Town Centre	1.2	1.3	1.2	1.3	1.2

TABLE 6.8: Mean Average Displacement error with quantized annotated head pose and with real head pose estimator.

6.4.4 Qualitative results

Besides these quantitative results and ablation studies, we report a qualitative illustration of our predictions in Fig. 6.4. Along with the proposed approach, we also show trajectories predicted with LTA [PESV09] and SF [YBOB11]. Notably, our model is able to better forecast trajectories with highly non-linear avoidance turns, such as to avoid static (3rd row, 3rd column) and moving objects (3rd row, 2nd column), as well as in case a person has to avoid collision with other pedestrians in the scene (1st, 2nd and 4th rows).

6.5 Data Driven Approaches for trajectory forecasting

Anticipating the trajectories that could occur in the future is important for several reasons: in computer vision, path forecasting helps the dynamics modeling for target tracking [PESV09; RSAS16; SAS17; YBOB11] and behavior understanding [AGRRF+16; KZBH12; LK16; MHLK17; RSAS16]; in robotics, autonomous systems should plan routes that will avoid collisions and be respectful of the human proxemics [DRS11; Hal66; KKSB12; MHB16; TK10; ZRGMP+09]. Recently, path forecasting has benefited from the introduction of Long Short Term Memory (LSTM) architectures [**ijcai2017-386**; AGRRF+16; HS97; SDZLZ16; SYMHD17; VS17].

All of these approaches use exclusively the (x, y) position coordinates for the prediction, forgetting that humans act and react using their senses to explore the environment, in particular, through the visual information conveyed by the gaze and inferred by the head pose [Cv80; CO12; DR12; FUCH15; FUY15; FWK11; IC01; PV03; RR06a; SFYW99; VCDPL13]. In particular, [Cv80; DR12; FUCH15; FUY15; FWK11; IC01; PV03; VCDPL13] found that the head pose correlates to the person destination and pathway: these findings are also supported by a statistical analysis presented in our paper (Sec. 7.3.1).

For the first time this work considers the head pose, jointly with the positional information, as a cue to perform forecasting. In particular, tracklets (sequences of (x, y) coordinates) and *vislets*, that is, reference points indicating the head pan orientation, are the input of the novel MiXing LSTM (MX-LSTM), an LSTM-based model that learns how tracklet and vislet streams are related, mixing them together in the LSTM hidden state recursion by means of cross-stream full covariance matrices, optimized during backpropagation.

MX-LSTM is able to encode how movements of the head and the people dynamics are connected. For example, it captures the fact that rotating the head towards a particular direction may anticipate a trajectory drifting with an acceleration (as in the case of a person leaving a group after a conversation). This happens thanks to a novel optimization of the LSTM parameters using a Gaussian full covariance through an unconstrained log-Cholesky parameterization in the backpropagation, securing positive semidefinite matrices. To the best of our knowledge, this is the first time Gaussian distributions with covariance matrices of order higher than two are optimized in LSTMs.

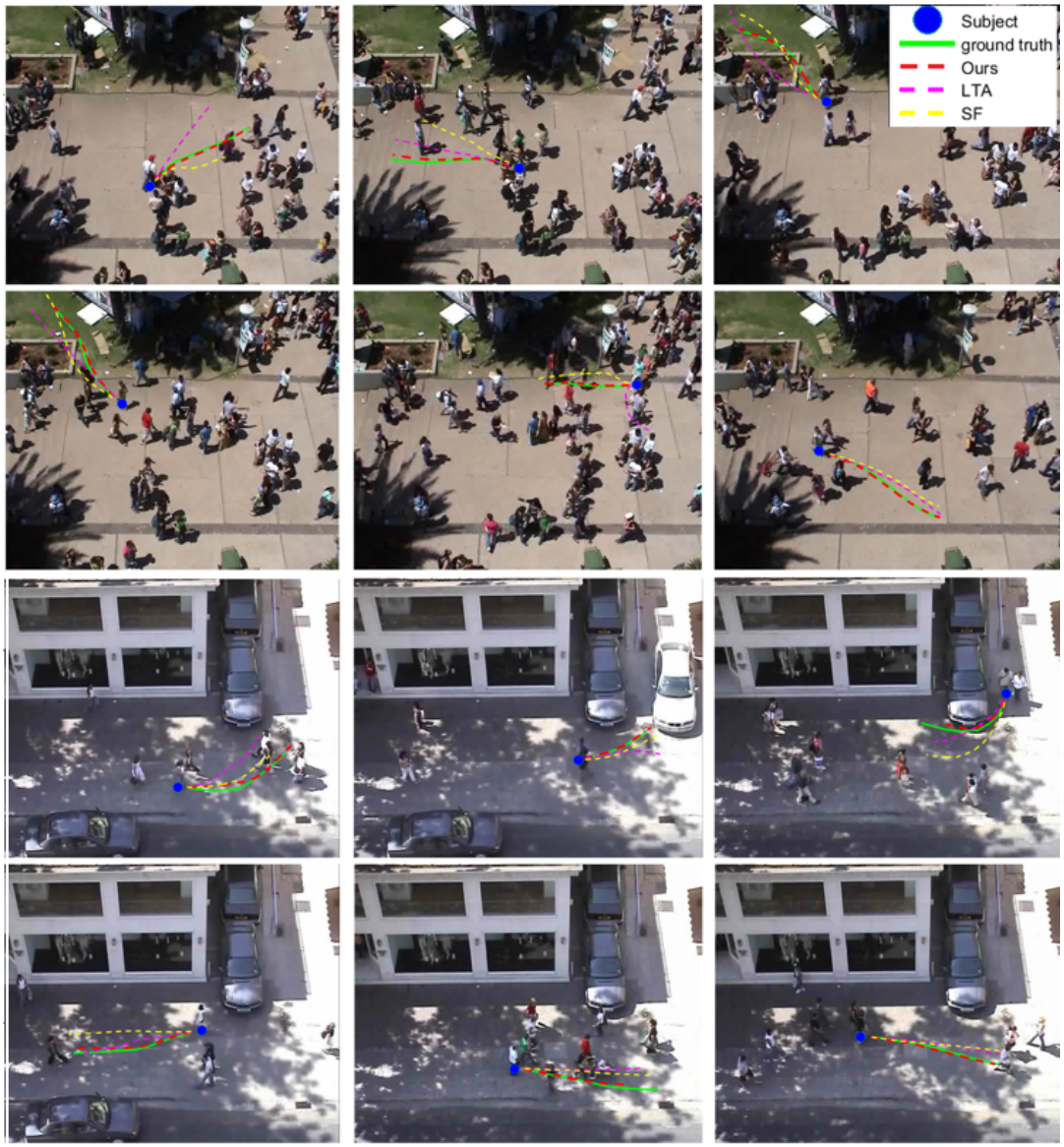


FIGURE 6.3: Examples of predicted trajectories on UCY (first two rows) and Zara01, Zara02 (last two rows). Our proposed model is very precise in the prediction of highly non-linear trajectories, where the other approaches such as LTA [PESV09] and SF [YBOB11] are less accurate due to the fixed destination points. In particular, our method is able to easily capture short term deviations from the desired path.

Vislet information is also used to build a scene context, i.e. where are the other people and how they are moving, by a shared state pooling as in [AGRRF+16; VS17], that here is further improved using the head pose by discarding the people that an individual cannot see.

As a by-product, MX-LSTM predicts head orientations too, allowing to reason where people will most probably look at, providing a fine grained level of long-term prediction never reached so far in crowded scenarios.

Adopting standard protocols for trajectory forecasting [AGRRF+16; LCL07; PESV09]

FIGURE 6.4: Illustration of common failure cases of trajectory forecasting. Acceleration, deceleration and static groups are common failure cases across all approaches.

and using head poses information given by a standard head pose estimator [HTGDC17], MX-LSTM defines the new state-of-the-art both in the UCY sequences (Zara01, Zara02 and UCY) and in the TownCentre dataset. In particular, MX-LSTM has the ability to forecast people when they are moving slowly, the Achille's heel of all the other approaches proposed so far.

As main contributions, in this work we show:

- We show that trajectory forecasting can be dramatically ameliorated by considering head pose estimates;
- We propose a novel LSTM architecture, MX-LSTM, which exploits positional (tracklets) and orientational (vislets) information thanks to an optimization of d-variate Gaussian parameters including full covariances with $d > 2$;
- We motivate the need for MX-LSTM showing that head poses are related with the trajectories, even at low velocities, where most of the forecasting approaches fail;
- We define a novel type of social pooling, in the sense of [AGRRF+16; VS17], by exploiting the vislet information;

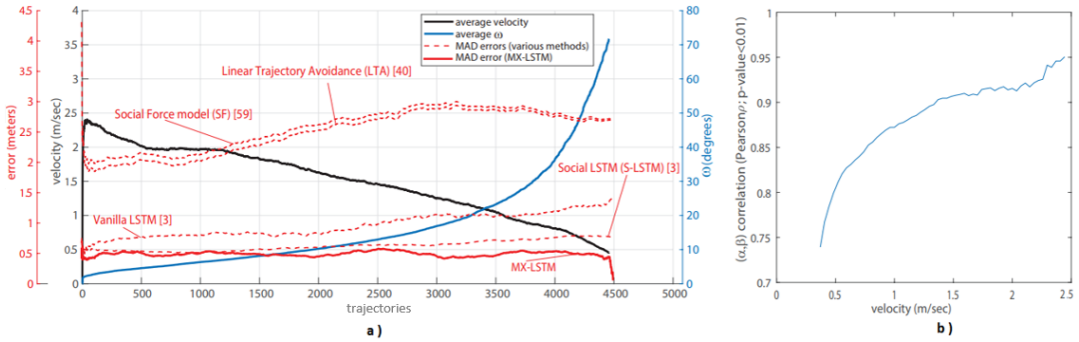


FIGURE 6.5: Motivating the MX-LSTM: a) analysis between the angle discrepancy ω between head pose and movement, the pedestrain smoothed velocity and the average errors of different approaches on the UCY sequence [LCL07]; b) correlation between movement angle β and head orientation angle α when the velocity is varying (better in color).

- Thanks to MX-LSTM, we define state-of-the-art forecasting results on different datasets;
- We present MX-LSTM results of head pose forecasting, showing new long-term behavior analysis capabilities.

6.5.1 Motivation of MX-LSTM

Intuitively, the head pose of person is a cue for the direction in which she/he moves. However, the literature in trajectory forecasting lacks a quantitative study on the importance of the head pose. Here we examine the common forecasting datasets to study the relationship between the head pose and motion directions. In particular, we focus on the UCY dataset [LCL07], composed by the Zara01, Zara02 and UCY sequences, which provides the annotations for the pan angle of the head pose of all the pedestrians. We also consider the Town Center dataset [BR11], where we have manually annotated the head pose, using the same annotation protocol as in [LCL07].

In this section, with specific reference to Figure 6.5, we present the preliminary analysis and observations, which have motivated the design of our MX-LSTM. We would specifically refer to the UCY video sequence (but similar observations applied to all others).

1) People watch their steps. We show this fact by plotting in Fig. 6.5a the angular discrepancy ω (*blue curve*), between the head pose α and the person motion angle β, against the velocity (*black curve*), intended as the modulus of the motion vector $\overrightarrow{x_{t+1} - x_t}$.

In more details, we have computed the average angular discrepancy ω for each of the people trajectories of the UCY video sequence (for each trajectory, we average ω across all frames where it occurs). In Fig. 6.5a, we have then arranged the trajectories in ascending order (the x axis) according to their average discrepancy angle ω (the *blue* y-axis on the sub-figure right side, marked as "ω"). (We illustrate ω graphically in Fig. 6.6c.) For each trajectory we have then plotted the corresponding average speed (*black curve*), as measured on the black y-axis marked as "velocity". (We disregard those frames where the average speed of person movement is below 0.45m/sec, since those people do not essentially move and their motion angle β can hardly be determined.)

As it shows from Fig. 6.5a, 75% of the people only turn their head by 20°. They watch therefore their steps, especially at higher speeds.

2) Head pose and movements are (statistically) correlated. On Fig. 6.5a, we report the velocity curve (black solid line and axis). To plot this curve, we order all the trajectories with respect to the average speed of each individual. First of all, notice that the ω and the

pedestrian speed are inversely proportional: the alignment between the head pose and the direction of movement is higher when the speed is higher; when the person slows down the head pose is dramatically misaligned. Secondly, the relation is statistically significant: we consider the Pearson circular correlation coefficient [JS01] between the angles α_t and β_t. Overall, the correlation is 0.83 (p-value$<$ 0.01), computed for all the frames of the sequences considered for Fig. 6.5. The plot in Fig. 6.5b elaborates that the correlation is lower at low velocities, where the discrepancy between the α_t and β_t angles is typically higher.

One of the challenges here, is to investigate whether the dynamic discrepancy between the head pose angle α_t and movement direction β_t at different speeds of the human motion can be learned by our proposed MX-LSTM to improve the forecasting. Moreover, MX-LSTM should learn how these relations evolve in time, which has not been investigated yet. In fact, prior work has only addressed single frames.

3) Forecasting is difficult for pedestrians at low speeds. In Fig. 6.5a (red lines and red axis), we compare the Mean Average Displacement (MAD) error [PESV09] of the following approaches: SF [YBOB11], LTA [TK10], vanilla LSTM and Social LSTM [AGRRF+16], against our proposed MX-LSTM approach (solid red curve). We notice that lower velocities correspond generally to higher forecasting errors. When people walk slowly, their behavior becomes less predictable, not only due to physical reasons (less inertia), but also behavioral (people walking slowly are usually involved in secondary activities, such as looking around or chatting with others). By contrast, our proposed approach MX-LSTM (solid red curve) performs well even at lower velocities, since it makes use of the evidence from the head pose. MX-LSTM approaches an error close to zero for the nearly static people, as it should ideally be (more details in Sec. 6.5.7).

Summarizing, the head pose is correlated with the movement. When people move fast, this correlation is stronger and their head pose is largely aligned with the direction of motion. However, when people move slowly, the correlation is weaker (but still significant), and the head pose is drastically misaligned with the movement. This results in higher prediction errors for most state-of-the-art approaches of trajectory forecasting. These facts justify and motivate our objective with the MX-LSTM, to capture the head pose information jointly with the movement and use it for a better and more uniform trajectory forecasting, for people moving at both lower or higher speeds.

6.5.2 Proposed Approach

In this section, we present *MX-LSTM*. The model may jointly forecast individuals' locations and pose by leveraging the information about the recent history of head positions (*tracklets*) and orientations (*vislets*). We first define the concepts of tracklets and vislets (Sec. 6.5.3); then we describe our proposed formulation of social pooling based on visual frustum of attention (Sec. 6.5.4); finally, we report details about the LSTM formulation (Sec. 6.5.5) and model training by optimizing the multidimensional co-variance matrices (Sec. 6.5.6).

6.5.3 Tracklets and vislets

We define as *tracklet* the list of consecutive locations on the ground plane visited by an individual during the last time steps. Formally, the tracklet associated with the i-th subject at time T is $\{\mathbf{x}_t^{(i)}\}_{t=1,\dots,T}$, where $\mathbf{x}_t^{(i)} = (x, y) \in \mathcal{R}^2$.

Similarly, a *vislet* is the list of anchor points located at a fixed distance r from the subject, aligned with its head head orientation. Thus, for subject i at time T the vislet is $\{\mathbf{a}_t^{(i)}\}_{t=1,\dots,T}$, with $\mathbf{a}_t^{(i)} = (x_t^{(i)} + \cos\alpha_t^{(i)}, y_t^{(i)} + \sin\alpha_t^{(i)}) \in \mathcal{R}^2$ see Fig. 6.6a.

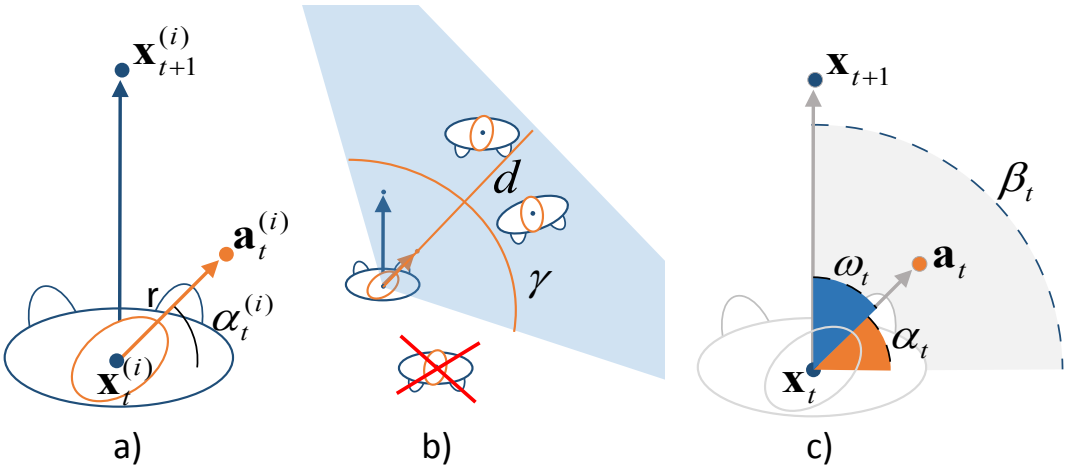


FIGURE 6.6: A graphical interpretation of tracklets and viselets. a) tracklets $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_{t+1}^{(i)}$ and vislet anchor point $\mathbf{a}_t^{(i)}$; b) Social pooling leveraging the Visual Frustum of Attention; c) angles for the correlation analysis.

In theory, one could encode the head orientation by means of the pan angle at each time step. We prefer to use anchor points instead, which gives several benefits. The main advantage of using vislets instead of encoding the head orientation directly with the pan angle, is that this formulation implicitly solve all the issues generated by the discontinuity between 360° and 0°. Moreover, vislets and tracklets have very similar representations, which is very convenient for modeling the interplay of these two components in the MX-LSTM structure. Please note that the distance r is irrelevant, as long as it is a constant value; in this work we set it at 0.5m for the sake of visualization.

Our method relies on a set of location and head pose observations to predict tracklets and vislets for the following estimation period. In particular, MX-LSTM mixes together the two streams to understand their relationship, providing a joint prediction. Accordingly to the trajectory forecasting literature [AGRRF+16; TK10; YBOB11], we consider these observations as provided by an oracle, *i.e.* given by an annotator. To directly compare our approach with the other recent ones, we provide experiments where the past head poses are estimated by a real "static" head pose estimator; in this way, MX-LSTM will require no additional effort in annotation with respect to former approaches.

We instantiate an LSTM model for each individual by using two separate embedding functions for tracklets (6.1) and vislets (6.2):

$$\begin{aligned} \mathbf{e}_t^{(x,i)} &= \phi\left(\mathbf{x}_t^{(i)}, \mathbf{W}_x\right) \\ \mathbf{e}_t^{(a,i)} &= \phi\left(\mathbf{a}_t^{(i)}, \mathbf{W}_a\right) \end{aligned}$$

where the embedding function ϕ is the linear projection, via the embedding weigths $\mathbf{W}_{(\cdot)}$, into a D-dimensional vector, with D the dimension of the hidden space. This is followed by a ReLU activation function.

6.5.4 VFoA social pooling

The concept of social pooling was first introduced by [AGRRF+16] as an effective way to capture (and embed into an LSTM model) how people move in a crowded space to avoid collisions. In its original form, it is an isotropic area of interest surrounding the target individual. The LSTM hidden variables of the people within the area of interest are pooled, *i.e.*

collected to account for the human-human interaction. This formulation implicitly assumes that a person's trajectory is affected not only by the behaviour of people walking in front of him/her, but also by people behind him/her back as also illustrated in Fig 6.7. in this work we refine this model by exploiting vislet information, building on the concept of View Frustum of Attention (VFoA), that is a region where the attention of a person is focused, according to its gaze direction. We propose to model the VFoA as a circular sector originating in the head position ($\mathbf{x}_t^{(i)}$), aligned with the head pose (*i.e.* towards the anchor point $\mathbf{a}_t^{(i)}$), with a aperture angle γ; to account for the limitations of human vision in focusing on very far ahead objects, we limit the region with a maximum distance d. We learned both γ and d parameters at training time by cross-validation on the training partition of the TownCentre dataset. A graphical interpretation of the VFoA is provided in Fig 6.6(b).

Formally, we define an area of interest as the squared region centered at the pedestrian location with size $2d \times 2d$; this area is then divided in a uniform grid of $N_o \times N_o$ cells. Our VFoA social pooling is a $N_o \times N_o \times D$ tensor $\mathbf{H}$ defined as follows:

$$\mathbf{H}_t^{(i)}(m,n,:) = \sum_{j \in \text{VFoA}_i} \mathbf{h}_{t-1}^{(j)}, \tag{6.3}$$

where the m and n indices run over the $N_o \times N_o$ grid and the condition $j \in \text{VFoA}_i$ is satisfied when the subject j is in the VFoA of subject i. The pooling vector is then embedded into a D-dimensional vector by

$$\mathbf{e}_t^{(H,i)} = \phi(\mathbf{H}_t^{(i)}, \mathbf{W}_H). \tag{6.4}$$

6.5.5 LSTM recursion

The MX-LSTM recursion equation is:

$$\mathbf{h}_t^{(i)} = LSTM\left(\mathbf{h}_{t-1}^{(i)}, \mathbf{e}_t^{(x,i)}, \mathbf{e}_t^{(a,i)}, \mathbf{e}_t^{(H,i)}, \mathbf{W}_{\text{LSTM}}\right).$$

The hidden state of the LSTM model projects onto the four dimensional space, representing the Gaussian multi-variate distribution $\mathcal{N}(\text{-}_t^{(i)}, \mathbf{\Sigma}_t^{(i)})$, as follows:

$$[\text{-}_t^{(i)}, \hat{\mathbf{\Sigma}}_t^{(i)}] = \mathbf{W}_o \mathbf{h}_{t-1}^{(i)},$$

where $\text{-}_t^{(i)} = [\mu_t^{(x,i)}, \mu_t^{(y,i)}, \mu_t^{(a_x,i)}, \mu_t^{(a_y,i)}]$, $\mathbf{\Sigma}_t^{(i)}$ contains the covariances among the (x,y) coordinate distributions of the tracklets and the vislets, and $\hat{\mathbf{\Sigma}}_t^{(i)}$ is its vectorized version. The distribution is then sampled to generate the joint prediction of tracklets and vislet points $[\hat{\mathbf{x}}_t, \hat{\mathbf{a}}_t]$, allowing us to simultaneously forecast trajectries and head poses.

At training time, we estimate the weights of the LSTM by minimizing the multivariate Gaussian log-likelihood for the each trajectory. The loss function is

$$L^i(\mathbf{W}_x, \mathbf{W}_a, \mathbf{W}_H, \mathbf{W}_{\text{LSTM}}, \mathbf{W}_o) \quad = \\ -\sum_{T_{obs}+1}^{T_{pred}} log\left(P([\mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}], \text{-}_t^{(i)}, \mathbf{\Sigma}_t^{(i)})\right),$$

where T_{obs} is the last frame of the observation period, while $T_{obs}+1, \ldots, T_{pred}$ are the time frames for which we provide a prediction. The loss of Eq. (6.5) is minimized over all the training sequences. To prevent overfitting, we additionally include an l_2 regularization term.

6.5.6 MX-LSTM optimization

As shown in Eq. (6.5), the optimization procedure provides the weight matrices of the MX-LSTM, which in turn produces the set of Gaussian parameters, including the full covariance $\mathbf{\Sigma}$. The latter is needed to enforce the LSTM in encoding the relations among the (x, y) coordinate distributions of tracklets and vislets, which we already discussed in Sec. 6.5.3. In principle, one may have simply captured the correlation between the walking direction and head pose in order to model drifts in the trajectory, but we are interested in letting the MX-LSTM analyze also how the head pose (pan angle) influences the length of the spatial step, that is the velocity. In other words, we want the MX-LSTM to be able to capture whether a particular head pose dynamics could accelerate or slow down the motion, thus letting the machine forecast the joint behavior.

The estimation of a full covariance matrix as the result of an optimization procedure over a generic objective function, like the log-likelihood of (6.5), is a difficult numerical problem [PB96]. The main reason is that one must guarantee that the resulting estimate is a proper covariance matrix, *i.e.* a positive semi-definite (p.s.d.) matrix. For this reason, LSTMs with log-likelihood loss functions over Gaussian distributions have been restricted so far to two dimensions, using a simple Gaussian [AGRRF+16], or mixture of Gaussian distributions. The 2×2 covariance matrices have been obtained by optimizing the scalar correlation index $\rho_{x,y}$, which becomes the covariance term of $\mathbf{\Sigma}$ with $\sigma_{x,y} = \rho_{x,y}\sigma_x\sigma_y$ [Gra13].

In case of higher dimensional problems, pairwise correlation terms cannot be optimized for building $\mathbf{\Sigma}$, since the optimization process for each correlation term is independent from each other. At the same time, the positive-definiteness is a simultaneous constraint on multiple variables [Pou11]. In practice, if we consider three variables x, y and z, learning $\rho_{x,y}$ and $\rho_{x,z}$ are two independent procedures, despite that they act on the common distribution over x. This lacks of coordination generates matrices far from being p.s.d. and thus requiring a further correction procedure, It usually consists of projecting the estimated matrix into the closest p.s.d. matrix based on a cost function of the Frobenious norm [BX05; Hig88]. This procedure is very expensive [PB96], and difficult to be embedded into the LSTM optimization process [DS96], where nonlinearities due to the embedding weights make the analytical derivation hard to formulate. So far, there is not any LSTM loss that involved full covariances of dimension higher than 2.

Our solution involves unconstrained optimization; we use an appropriate Cholesky parameterization of the matrix to be learned that enforces the positive semi-definite constraint, dramatically improving the convergence properties of the optimization algorithm [Pou11]. Let us consider $\mathbf{\Sigma}$ a definite positive $n \times n$ (in our case, $n = 4$) covariance matrix. Since $\mathbf{\Sigma}$ is symmetric by definition, only $n(n+1)/2$ parameters are required to represent it. The Choleski factorization is given by:

$$\mathbf{\Sigma} = \mathbf{L}^T\mathbf{L}, \tag{6.6}$$

where $\mathbf{L}$ is a $n \times n$ upper triangular matrix. The optimization process focuses on finding the $n(n+1)/2$ distinct scalar values for $\mathbf{L}$, which we then solve for the covariance, as for Eq. (6.6). The main problem with the Cholesky factorization is non-uniqueness: any matrix obtained by multiplying a subset of the rows of $\mathbf{L}$ by -1 is still a valid solution. As a consequence, non-uniqueness makes the problem ill-posed and hinders optimization convergence. The simplest way to enforce the matrix $\mathbf{L}$ to be unique is to add the constraint that all the diagonal elements must be positive. To this end, the Log-Cholesky parameterization [Pou11] assumes that the values found by the optimizer of the main covariance diagonal are the log

FIGURE 6.7: VFOA pooling: For a given subject, he will try to avoid collision with the people who are inside his view frustum (blue circle). Others (red circle), will not influence his trajectory as they are no in his view frustum.

of the values of $\mathbf{L}$. Formally, the values found by the optimizer can be written as:

$$\grave{}_L = \begin{bmatrix} \log l_{1,1} & l_{1,2} & l_{1,3} & l_{1,4} \\ 0 & \log l_{2,2} & l_{2,3} & l_{2,4} \\ 0 & 0 & \log l_{3,3} & l_{3,4} \\ 0 & 0 & 0 & \log l_{4,4} \end{bmatrix}.$$

In practice, after the estimation of $\mathbf{W}_x$, $\mathbf{W}_a$, $\mathbf{W}_H$, $\mathbf{W}_{\text{LSTM}}$, $\mathbf{W}_o$ parameters, the values of $\grave{}_L$ are extracted by

$$[\bar{}_t^{(i)}, \hat{\grave{}}_{Lt}^{(i)}] = \mathbf{W}_o \mathbf{h}_{t-1}^{(i)},$$

where $\hat{\grave{}}_L$ is the vectorized version of $\grave{}_L$. Then, the diagonal values of $\grave{}_L$ are exponentiated to form $\mathbf{L}$ and obtaining $\mathbf{\Sigma}$ through Eq. (6.6).

6.5.7 Experiments

To validate the proposed approach we perform both qualitative and quantitative evaluations. We report experiments on two public datasets, namely *UCY* [LCL07] and *TownCentre* [BR11] datasets. We compare our model with one baseline, *i.e.* a standard LSTM model that only accounts for pedestrian positions (Vanilla LSTM), and four state-of-the-art approaches: Social Force model (SF) [YBOB11], Linear Trajectory Avoidance (LTA) [PESV09], Social LSTM (S-LSTM) [AGRRF+16] and Social GAN [GJFSA18a]. Here we also investigate three variations of the MX-LSTM model to capture the net contributions of the different parts that characterize our approach. Moreover, we investigate the effect of changing the observation period and the forecasting horizon, illustrating how head pose plays a pivotal

Metric	Dataset	SF [YBOB11]	LTA [PESV09]	Vanilla LSTM [AGRRF+16]	Social LSTM [AGRRF+16]	Social GAN [GJFSA18a]	MX-LSTM	*MX-LSTM-HPE*	Individual MX-LSTM	NoFrustum MX-LSTM	BD- MX-LSTM
MAD	Zara01	2.88	2.74	0.90	0.68	*0.48*	**0.59**	*0.66*	0.63	0.63	0.60
	Zara02	2.32	2.23	1.09	0.63	*0.44*	**0.35**	*0.37*	0.72	0.36	0.41
	UCY	2.57	2.49	0.67	0.62	*0.65*	**0.49**	*0.55*	0.53	0.51	0.54
	TownCenter	9.35	9.14	4.62	1.96	*1.60*	**1.15**	*1.21*	2.09	1.70	1.40
FAD	Zara01	5.55	5.55	1.85	1.53	*1.04*	**1.31**	*1.43*	1.37	1.40	1.51
	Zara02	4.35	4.35	2.15	1.43	*0.95*	**0.79**	*0.82*	1.56	0.84	1.00
	UCY	4.62	4.66	1.39	1.40	*1.36*	**1.12**	*1.20*	1.16	1.15	1.23
	TownCenter	16.01	16.08	8.26	3.96	*3.50*	**2.30**	*2.38*	4.00	3.40	2.90

TABLE 6.9: Mean and Final Average Displacement errors (in meters) for all the methods on all the datasets. The first 6 columns are the comparative methods and our proposed model trained and tested with GT annotations. MX-LSTM-HPE is our model tested with the output of a real head pose estimator [HTGDC17]. The last 3 columns are variations of our approach trained and tested on GT annotations.

role for the long term forecasting. Lastly, we analyze whether one can substitute the ground-truth head pose information with more accessible proxies, such as the pace direction or head pose estimates, as provided by a detector. On a qualitative evaluation, we show the interplay between tracklets and vislets that the MX-LSTM has learnt.

6.5.8 Implementation details

We implemented the MX-LSTM model and all models of the ablation study in Tensorflow. All models have been trained with learning rate of 0.005 along with the RMS-prop optimizer. We set the embedding dimension for spatial coordinates and vislets to 64 and the hidden state dimension is D = 128. We compute the social pooling on a grid of 32×32 cells (6.3) The view frustum aperture angle has been cross-validated on the training partition of the TownCentre and kept fixed for the remaining trials ($\gamma = 40°$), while the depth d is simply bounded by the social pooling grid. Training and testing has been accomplished with a GPU NVIDIA GTX-1080 for all evaluations.

6.5.9 Evaluation Protocol

We report experiments on two public datasets, namely *UCY* [LCL07] and *TownCentre* [BR11] datasets.

The evaluation protocol follows the standard procedure for trajectory forecasting that is used in the literature [PESV09; AGRRF+16]. We first downsample the videos at 0.4fps, then we observe tracklets and vislets for 8 frames, and we predict both locations and head poses for the following 12 time steps. The observation period is $3.2s$ and the forecasting horizon is $4.8s$. Experiments with different time horizons are reported in the ablation study (Sec. 6.5.11). According to the standard protocol, we use annotations during the observation period. Since we use additional information with respect to most of the related approaches (*i.e.* head poses), we perform an evaluation with the output of a real head pose estimator as well (Sec. 6.5.11).

For the three UCY sequences we train three models, where we use two sequences for training and the remaining for testing. For the TownCentre dataset, the model has been trained and tested on the provided data splits.

Regarding the evaluation metrics of the trajectory forecasting, we consider the *Mean Average Displacement* (MAD) error, *i.e.* the average Euclidean distance between all the predicted and ground-truth pedestrian locations. The *Final Average Displacement* (FAD) error, *i.e.* the Euclidean distance between the last predicted location of each trajectory and the corresponding manually annotated point, is employed as well. Lastly, we evaluate the performance of the head pose predictions in terms of mean angular error e_α, which is the mean absolute difference between the estimated pose and the annotated ground truth.

6.5.10 Comparison with Prior Art

We compare our model against a baseline Vanilla LSTM model, which only uses pedestrian positions, and four state-of-the-art approaches: Social Force model (SF) [YBOB11], Linear Trajectory Avoidance (LTA) [PESV09], Social LSTM (S-LSTM) [AGRRF+16] and Social GAN [GJFSA18a].

Note that the Social GAN [GJFSA18a] uses ground-truth trajectories during the prediction interval: At test time, the Social GAN [GJFSA18a] model predicts 20 trajectories and uses the L_2 distance w.r.t. the ground-truth trajectory to select the best one. Although this protocol makes the comparison with all other approach unfair, we include it in the results for the sake of completeness.

Comparative results are reported in Table 6.9. The MX-LSTM outperforms the state-of-the-art methods across all sequences on both metrics, with an average improvement of 23.3% over the second best performer, Social GAN. The highest relative gain is achieved in the UCY sequence and TownCentre dataset, where we achieve a MAD error of 0.49 and 1.15 respectively, improving on Social GAN by 24% and 28% respectively. We explain the larger relative improvement by the increased difficulty of the complex non-linear people paths, in which case the visual attention turns out an important cue. In UCY and TownCenter, people stand in conversational groups, others walk by closely, while some of them slow down to look at the shop windows. We provide quantitative examples of these complex motions in Fig. 6.5.

Note that some of the evaluated methods require additional input data: both SF and LTA require the destination point of each individual, while SF additionally requires the social group annotations. Ours uses the manually labelled (ground-truth) head poses, which are provided to the algorithm (only) in the observation period (before the forecast). We discuss in the next subsection whether this manual annotation is really needed.

Effect of head pose estimator

Here we analyze the effect on performance, at inference time, of adopting a head pose estimation algorithm [HTGDC17] during the observation period (prior to forecasting), instead of the ground-truth head poses.

We automatically estimate the head bounding box given the feet positions on the floor plane, assuming an average person being 1.80m tall. Then, we apply the head pose estimator of [HTGDC17] that provides continuous angles for the pan orientation. At inference time, this data is used as input to this variant, which we name "MX-LSTM-HPE".

Results in Table 6.9 illustrate that the performance of MX-LSTM-HPE is in average 9% worse than MX-LSTM. The importance of the head pose estimate quality for forecasting is therefore notable, which makes future research on head pose an indispensable requirement. Note from Table 6.9 that the results of MX-LSTM-HPE (MAD and FAD, across all sequences) are still better than any other competing approach.

6.5.11 Ablation Study

We analyse the net contribution of different parts of the proposed approach by investigating three variations of our model: namely *Block-Diagonal*, *NoFrustum* and *Individual* MX-LSTM.

Block-Diagonal MX-LSTM (BD-MX-LSTM): This studies the importance of estimating full covariances to understand the interplay between tracklets and vislets, rather than modelling each of them as a separate probability distribution. Essentially, instead of learning the 4×4 full covariance matrix $\mathbf{\Sigma}$, BD-MX-LSTM estimates two separate bidimensional

covariances $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_a$ for the trajectory and the vislet modeling, thus neglecting the cross-stream covariance. Each 2×2 covariance is estimated employing two variances σ_1, σ_2 and a correlation terms ρ as presented in [Gra13].

NoFrustum MX-LSTM: this variant reduces MX-LSTM to the social pooling of [AGRRF+16], i.e. pooling for hidden states $\{\mathbf{h}_t^j\}$ from the entire area around each individual. NoFrustum MX-LSTM neglects the visual frustum of attention and does not select the people to pool from based on it. Also people behind the person would therefore influence the next step forecasting.

Individual MX-LSTM: in this case, no social pooling is taken into account. In more detail, the embedding operation of Eq. (6.4) is removed, and the weight matrix $\mathbf{W}_H$ vanishes. In practice, this variant learns independent models for each person, each one considering the tracklet and vislet points.

he last three columns of Table 6.9 report numerical results for the three MX-LSTM variants. The main facts that emerge are: 1) the highest variations are with the Zara02 sequence, where MX-LSTM doubles the performances of the worst approach (Individual MX-LSTM); 2) the worst performing is in general Individual MX-LSTM, showing that social reasoning is indeed needed; 3) social reasoning is systematically improved with the help of the vislet-based view-frustum; 4) full covariance estimation has a role in pushing down the error which is already small with the adoption of vislets.

Summarizing the results so far, having vislets as input allows to definitely increase the trajectory forecasting performance. Vislets should be used to understand social interactions with social pooling, by building a view frustum that tells which are the people currently observed by each individual. All of these features are effectively and efficiently implemented within MX-LSTM. Note in fact that the training time is not affected by whether social pooling is included or not.

Again, although the complete method always outperforms all the competitors, the highest improvement is on the TownCentre sequence. In our opinion this is due to the different level of complexity in the data, indeed most of the trajectories in UCY sequences are relatively linear, with poor social interactions, while in TownCentre there are many interactions, such as forming and splitting groups and crossing trajectories. For the same reason, this is the dataset where the introduction of the view frustum in the pooling of social interactions gives the highest benefits. By contrast, in all other sequences but Zara01, decoupling the covariance matrix into a block diagonal matrix neglecting the interplay of position and gaze (BD-MX-LSTM) leads to a sensitive increase in the prediction error; this proves the tight relation between the head orientation and the motion of an individual.

6.5.12 Head Pose Forecasting

Our MX-LSTM model also provides a forecast of the head pose of each individual at each frame, for the first time. We evaluate the performances of this estimation in terms of mean angular error e_α, i.e. the mean absolute difference between the estimated pose (angle $\alpha_{t,\cdot}$ in Fig. 6.6c) and the annotated ground truth. e_α expresses how much the direction in which an individual is looking at a particular time instant is different from the true one. This error measure is independent from the error in the predicted position. In other words, e_α measures the error in the gaze forecasting.

Table 6.10 reports numerical results of the static head pose estimator [LYO15b] (HPE), the proposed model fed with manually annotated head poses (MX-LSTM) and with the output of HPE (MX-LSTM-HPE) during the observation period. In all the cases our forecast output is comparable with the one of HPE, but in our case we do not use appearance cues – i.e. we do not look at the images at all. In the case of Zara01, the MX-LSTM is even better

Metric	HPE [HTGDC17]	MX-LSTM	MX-LSTM-HPE
Zara01	14.29	12.98	17.69
Zara02	20.02	20.55	21.92
UCY	19.90	21.36	24.37
TownCentre	25.08	26.48	28.55

TABLE 6.10: Mean angular error (in degrees) for the state-of-the-art head pose estimator [HTGDC17], and our model fed with manual annotations (MX-LSTM) and estimated values (MX-LSTM-HPE).

Dataset	Forecasting horizon	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**H = 12**	0.90	0.68	**0.59**	0.72
	H = 16	1.21	1.00	**0.87**	1.05
	H = 20	1.70	1.43	**1.21**	1.44
	H = 24	2.30	1.94	**1.55**	1.85
	H = 28	3.07	2.35	**1.92**	2.47
	H = 32	4.11	2.85	**2.40**	3.14
Zara 02	**H = 12**	1.09	0.63	**0.35**	0.63
	H = 16	1.62	0.90	**0.53**	1.09
	H = 20	2.19	1.24	**0.71**	1.43
	H = 24	2.75	1.59	**0.90**	1.83
	H = 28	3.31	2.00	**1.16**	2.25
	H = 32	3.86	2.41	**1.40**	2.67
UCY	**H = 12**	0.67	0.62	**0.49**	0.53
	H = 16	0.90	0.90	**0.70**	0.77
	H = 20	1.19	1.08	**0.95**	1.01
	H = 24	1.52	1.36	**1.22**	1.27
	H = 28	1.87	1.66	**1.50**	1.53
	H = 32	2.24	1.99	**1.80**	1.83

TABLE 6.11: Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 8 frames.

that the static prediction, which highlights the forecasting power of our model. In our opinion, this is due to the fact that in this sequence trajectories are mostly linear and that people are walking fast, with their heads mostly aligned with the direction of motion. When providing the MX-LSTM model with the estimations during the observation period, the angular error increases, as expected, but the error remains limited.

6.5.13 Time Horizon Effect

To investigate how MX-LSTM performs for longer time horizons we conduct an experimental evaluation where we increment the prediction interval from 12 (standard evaluation protocol) to 32 frames with a step size of 4, keeping the observation interval fixed at 8 frames. We evaluated approaches on UCY, Zara01 and Zara02, since most trajectories on TownCenter last less than 24 frames. We use MAD to report the error. As shown in Table 6.11, MX-LSTM is well capable of handling longer time horizons. MX-LSTM outperforms all other approaches on all prediction interval, which demonstrates its robustness. Based on these results, we argue that reasoning on the head pose becomes even more important when forecasting in the longer term. Overall, the ranking is preserved and MX-LSTM remains the best performer.

In order to understand how many frames are enough to learn a meaningful representation of the trajectory we varied the observation interval. Table 5 reports numerical results of

Dataset	Observation period	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**O = 1**	1.62	**0.89**	0.96	1.43
	O = 4	0.90	0.69	**0.64**	0.79
	O = 8	0.90	0.68	**0.59**	0.72
	O = 12	0.90	0.68	**0.59**	0.68
	O = 16	0.90	0.68	**0.59**	0.60
Zara 02	**O = 1**	1.65	1.13	**0.85**	1.35
	O = 4	1.17	0.74	**0.48**	0.84
	O = 8	1.09	0.63	**0.35**	0.63
	O = 12	1.01	0.63	**0.35**	0.63
	O = 16	0.99	0.63	**0.33**	0.62
UCY	**O = 1**	0.82	0.71	**0.62**	0.88
	O = 4	0.65	0.63	**0.49**	0.59
	O = 8	0.67	0.62	**0.49**	0.53
	O = 12	0.65	0.60	**0.48**	0.52
	O = 16	0.63	0.60	**0.48**	0.52

TABLE 6.12: Mean Average Displacement (MAD) error when changing the observation period. Forecasting horizon is kept constant at 12 frames.

Dataset	Prediction interval	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**Pred = 16**	1.25	1.05	**0.88**	0.90
	Pred = 20	1.27	1.46	**1.19**	1.26
	Pred = 24	1.78	1.88	**1.57**	1.64
	Pred = 28	2.39	2.37	**1.93**	2.01
	Pred = 32	3.09	3.00	**2.32**	2.57
Zara 02	**Pred = 16**	1.31	0.88	**0.49**	0.95
	Pred = 20	1.87	1.24	**0.67**	1.28
	Pred = 24	2.50	1.61	**0.87**	1.65
	Pred = 28	3.19	2.05	**1.11**	2.04
	Pred = 32	3.87	2.53	**1.35**	2.42
UCY	**Pred = 16**	1.02	0.80	**0.71**	0.72
	Pred = 20	1.42	1.06	**0.95**	1.01
	Pred = 24	1.87	1.34	**1.2**	1.40
	Pred = 28	2.37	1.67	**1.46**	1.50
	Pred = 32	2.92	2.21	**1.80**	1.90

TABLE 6.13: Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 16 frames.

an experiment where we kept the forecasting horizon fixed at 12 frames, and varied the observation period form 1 to 16 frames with the step size of 4 frames. An observation period of 1 frame means we try to predict trajectories based only on a static observation of the individual, with no previous history taken into account. Results prove that one frame is not enough for all the methods under analysis. Despite this, the ranking of different approaches is maintained throughout all the experiments, with the only exception of Zara01 sequence with O=1, where Social LSTM outperforms competitors. Interestingly, a rapid drop in error of about 30% is obtained by observing 4 frames instead of 1. Furthermore, 8 frames are enough for the approaches to learn the overall shape of the trajectory in order to predict for the next 12 frames, as the error drop from observing 8 frames to 16 frames is below 1%.

Finally, in order to understand in more depth how different methods perform for long term forecasting, we kept the observation interval constant at 16 frames and test increasing forecasting horizons. Table 6.13, further validates the fact that 8 frames are sufficient for the LSTM approach to learn the representation of the trajectory. MX-LSTM is still the best performer but the error drop from observing 8 to observing 16 frames is negligible in long

Dataset	MX-LSTM	MX-LSTM-HPE (Train and Test)	Pace-MX-LSTM
Zara01	0.59	0.68	0.69
Zara02	0.35	0.51	0.73
UCY	0.49	0.58	0.59
Town Centre	1.15	1.43	1.50

TABLE 6.14: MAD errors on the different datasets

term forecatsing as well. This effect speaks about the capability of LSTM-based approaches. The performance already starts to saturate at 8 frames and adding more information does not bring the expected gain. In our view, this highlights the temporal modelling as one of the performance bottlenecks, on the way to progress in the field.

6.5.14 Substitutes for Head Pose

In this experiment, we analyze the importance of the head pose and question whether one may substitute it with more accessible proxies, such as the direction of the people pace. In more details, we implement a Pace-MX-LSTM, which uses ground truth step directions instead of the head pose. Table 6.14 illustrates that having the step direction instead of the head pose downgrades the MX-LSTM, since positional data are already contained in the tracklet and the step direction can be extracted from the previous two positions. In fact, Pace-MX-LSTM gives consistently worse results.

In Table 6.14, we additionally illustrate the importance of having access to manually annotated head poses during training. to study this aspect, we implemented the MX-LSTM-HPE-Train and Test, where the head pose training data is given by a head-pose detector [HT-GDC17]. As expected, MX-LSTM-HPE-Train and Test underperforms MX-LSTM and MX-LSTM-HPE (MX-LSTM-HPE is still trainned on manually labelled head poses, but it adopts a head pose estimator at inference time). This is especially so on Zara02, where conversational groups make the head pose estimation noisy due to the many partial occlusions. Still, MX-LSTM-HPE-Train and Test remains comparable to prior state-of-the-art methods.

6.5.15 Qualitative Results

Fig. 6.8 shows qualitative results on the Zara02 dataset, which was found as the most difficult throughout the quantitative experiments. Fig. 6.8a presents MX-LSTM results: a group scenario is taken into account, with the attention focused on the girl in the bottom-left corner. In the left column, the green ground-truth prediction vislets show that the girl is havign a conversation with the group members, nearly not moving at all, while moving her head around. The magenta curve (Fig. 6.8a *left*) represents the S-LSTM output, predicting erroneously that the girl would leave the group. This error confirms the problem of competing methods in forecasting the motion of people slowly moving or static, as discussed in Sec. 6.5.1. In the central column of Fig. 6.8a, the observation sequence given to the MX-LSTM is shown in orange (almost static with oscillating vislets). The output prediction (yellow) shows oscillating vislets but no movement, confirming that the MX-LSTM has learnt this particular social behavior. If we provide the MX-LSTM with an artificial observation sequence with the annotated positions (real trajectory) but vislets oriented toward west (third column in Fig. 6.8a, orange arrows), where no people are present, the MX-LSTM predicts a trajectory slowly departing from the group (cyan trajectory and arrows).

The two rows of Fig. 6.8b analyze the Individual MX-LSTM, in which no social pooling is taken into account. Here pedestrians are not influenced by the surrounding people, and the forecast motion is only caused by the relationship between the tracklets and the vislets.

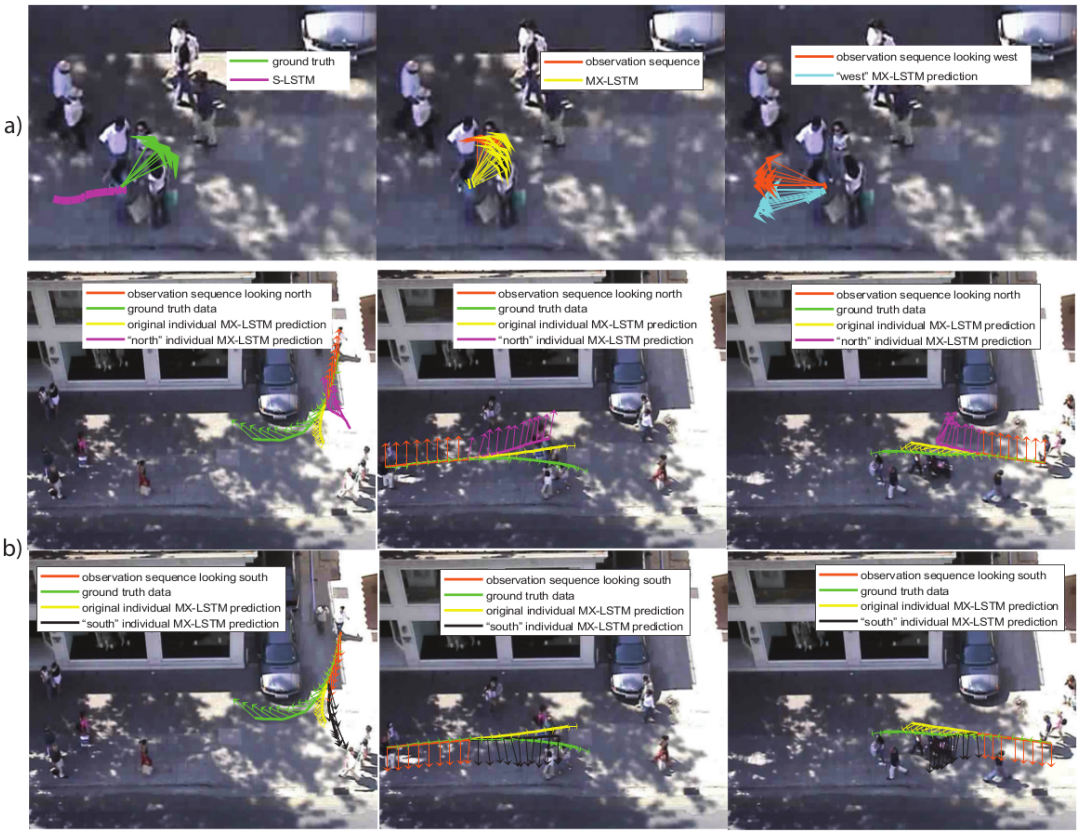


FIGURE 6.8: Qualitative results: a) MX-LSTM b) Ablation qualitative study on Individual MX-LSTM (better in color).

The first row in Fig. 6.8b shows three situations in which the vislets of the observation sequence are manually altered to point north (orange arrows), thus orthogonal to the person trajectory. In this case the Individual MX-LSTM predicts a decelerating trajectory drifting toward north (magenta trajectory and vislets), especially visible in the second and third rows. If the observation has the legit vislets (green arrows, barely visible since they are aligned with the trajectory), the resulting trajectory (yellow trajectory and vislets) has a different behavior, closer to the GT (green trajectory and vislets). Similarly, in the second row, we altered vislets to point to South. The prediction with the modified vislets is in black. The only difference is in the bottom left picture: here the observation vislets pointing south are in agreement with the movement, so that the resulting predicted trajectory is not decelerating as in the other cases, but accelerating toward south.

Chapter 7

Human-Centric Light Sensing and Estimation

7.1 Introduction

A modern lighting system should automatically calibrate itself (determine the type and position of lights), assess its own status (which lights are on and how dimmed), and allow for the creation or preservation of lighting patterns, e.g. after the sunset. The lighting patterns should be adjusted in a way, that is optimal for people actions and locality. As most of our activities hold within a given light pattern [FHSM79]. Moreover, light influences our perception of space [GV06], for example we expect to see a certain illumination pattern in a musical concert or a theater etc. The essence of such a system would be to deploy an *invisible light switch*, where the change in illumination is not perceived by the user.

Furthermore, idea of a smart lighting system, is to deploy a dynamic illumination pattern for a given activity, where the user have the sensation of "all-lit", while in reality the scene is optimally lit. In brief, this chapter discusses both fundamental research in computer vision and innovation transfer in smart lighting with a goal being at researching and developing novel autonomous tools using advanced computer vision and machine learning approaches that seamlessly integrates into smart lighting systems for indoor environments.

In this chapter, we propose a plan to create such an achievement, in light management systems, by enabling the understanding of the environment via long-term observation, that span days, weeks and even months, with a sensing device (i.e. RGB cameras or RGBD if including a depth sensor) for smart illumination and energy saving via an artificial intelligence (AI) processor (e.g. an algorithm to understand the scene and make decisions on lighting). More specifically in this Research and Development plan, top-view time-lapse images of the scene allow computer vision algorithms to understand it. In this work we try to estimate the human activities from RGB and RGBD images: in particular, recognize which and where activities occur in the environment, using technologies of detection and head pose estimation.

This chapter is a result of a joint work. Therefore, the experimental section is shared between two theses. The work related to detection and head pose estimation was primarily the contribution of this thesis alone.

7.2 Ego-light-perception

Any light management system that has to autonomously adjust the illumination of the environment has to be aware of two main factors: the human occupancy and their activity in the environment (human centric analysis) and the existing ambient illumination over time considering how is this influenced from the scene structure, the object materials and the light sources (scene composition analysis).

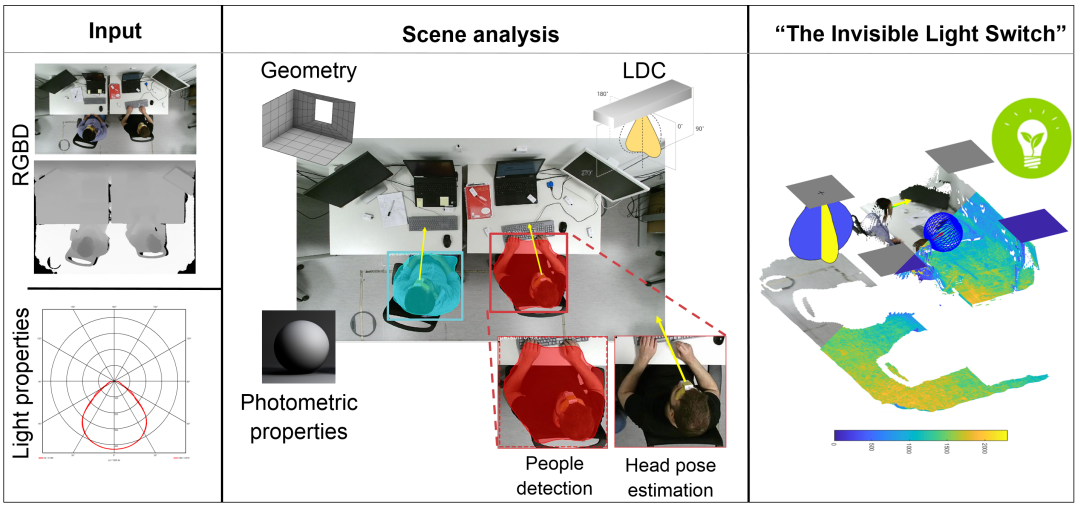


FIGURE 7.1: Overall pipeline of our system. We first acquire the RGBD input from the camera system *(left)* and together with the lighting system properties we use this information to create the Invisible Light Switch (ILS). That is, structuring the geometry of the scene, extracting the photometric properties of the material and applying a human centric analysis from where we detect the human presence in the scene and extract the possible head poses. Lastly we utilize the output of the scene analysis as the "Invisible Light Switch" application targeting a power saving framework.

Thus the Invisible Light Switch (ILS) is presented, as a smart lighting framework for dynamically adjusting the illumination level in an indoor environment. ILS takes into account the geometry of the scene, the presence of people and their light perception with the goals of maximizing the human comfort in terms of perceived light and, at the same time, with the lowest cost in terms of energy consumption. We do this by bringing together individual works into a unique pipeline as we show in Figure 7.1. The framework builds upon the light estimation system [THCGD19], which is capable of estimating the light in a given 3D point of a multi luminaire indoor environment. As we have shown, the presented radiosity model has been customized to take into account a realistic model of light propagation, outclassing even industrial software in the task.

We further enriched that model by including the human aspect, and showing how the interplay between the light estimation system and the human activity may lead to a consistent energy saving framework. The invisible light switch summarises the idea: an individual has the feeling of an environment which is globally illuminated, while in reality an automated light switch dims the luminaires in a way which is invisible to the users. This was possible by estimating the position of a person in the sensed environment, its head orientation, and understanding the light which is perceived by him. In fact, the lighting sensed by a human can be assumed as the light contained in a conic volume departing from the mean point connecting the human's eyes in the direction of the nose. Given this, it is possible to determine which luminaries could be switched off/dimmed down while maintaining the level of perceived light unchanged. The head pose is provided by detecting the person first and then estimating the head orientation. The former is carried out by means the state-of-the-art detector Mask R-CNN [**he2017mask**] with ResNet [**he2016deep**] as a backbone architecture, while head pose is done using Hasan's *et al*. method [HTGDC17].

7.2.1 People detection and head-pose estimation

We aim to detect people and estimate their head pose (their viewing angle). For the first task as we mentioned we adapted the Mask R-CNN [**he2017mask**] object detector, while for the second one the head pose estimator proposed in [HTGDC17].

The R-CNN [**he2017mask**] detector has the ResNet-101 [**he2016deep**] as a backbone architecture, trained on 80k images and 35k subset of evaluation images (trainval35k) of MS COCO dataset [LMBHP+14]. We fine-tuned the detector on our top-view dataset (see Sec. 7.3.1), adopting a specific training portion of the data. We randomly partitioned the data into training and testing set, keeping 70% of the data for training and 30% for testing. Since the top-view images are different from the frontal-view images of the COCO dataset [LMBHP+14], the fine-tuning had a crucial role. We adopted a similar procedure for training the head pose estimator as in [HTGDC17]. It is worth noting that the input for the head pose is the whole body detection bounding box: this is because [HTGDC17] has been specifically designed for managing small-sized head patches, exploiting the body as contextual cue for a better final head orientation classification. In particular, 4 and 8 classes related to angles have been taken into account.

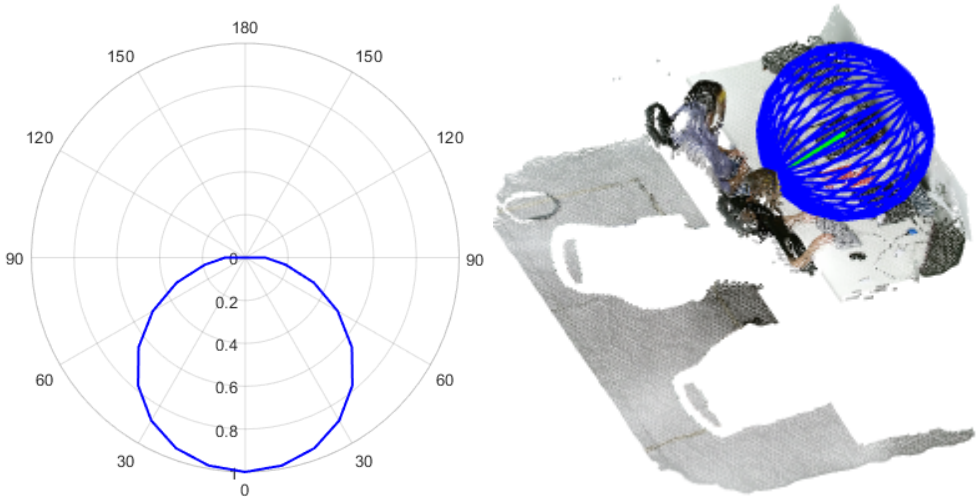
During testing time, a cascaded approach is followed, first by applying the people detector and then feeding the detected body bounding box as input into the head orientation module.

7.2.2 Spatial light estimation

To obtain an estimate of a dense spatial illumination map, we adapted our pipeline presentedin Chapter 4. As we have presented there we make use of a radiosity model [**cohen1993rri**] for estimating the spatial illumination over time by just using the input from an RGBD camera. Furthermore, we extract the information regarding the photometric properties of the material of the scene based on a photometric stereo baseline approach that is applied on the time-varying RGB images. This approach allows us to extract a scalar albedo at each pixel by using a set of images with different light sources that are switched on/off during the day. Having the light sources position and intensity, the scalar albedo under Lambertian assumptions, and the depth map from the sensor, our proposed method in Chapter 4 showed that it is possible to obtain a dense measurement of the light emitted by a 3D patch in the indoor environment. In order to provide more realistic estimates, we have shown how to model real lighting systems that, differently from point-like sources, emit light given a specific light distribution curve (LDC). The LDC is custom for each lighting system and their properties are considered to be known when estimating the light instensity. The proposed method shows that, even by accounting the non-linearities of LDC, it is possible to solve for the radiosity equation with Least Squares and so obtain a more reliable measure of the light intensity, which we evaluated by using point-to-point sensory equipment *aka*. luxmeters installed across the scene.

7.2.3 Gaze-gathered light modelling

Light measurements are practically made using a luxmeter sensor. This sensor measures the perceived light that is in function of the distance to the light, the orientation and other manufacturing characteristics. These properties are resumed by the Luxmeter Sensitivity Curve (LSC) as in Figure 7.2a. The LSC illustrates the perception characteristic of every luxmeter sensor which in this work we adopt in order to meet the measuring requirements of the collected ground truth data and to simulate the human light perception. We have chosen this solution because this is the standard de facto in the lighting industry and it provides satisfactory solutions when doing light commissioning [**ies2011commissioning**].



(A) (B)

FIGURE 7.2: Modeling of the Luxmeter Sensitivity Curve (LSC) as a human light perception model.

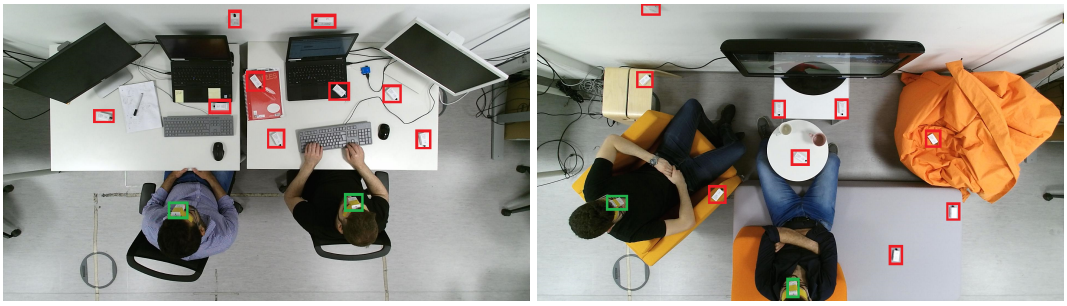
The key idea in this procedure is that, once we have detected a person in the image and estimated his head positioning and orientation as described in Sec. 7.2.1, we extract his posture in the 3D space by mapping the 2D image coordinates of his detected head to the corresponding depth information. Thereafter, once we have the positioning of the head in the 3D space as well as its orientation (where the person looks at), we estimate the light that arrives to his/her face (or to the luxmeter as in our case) by applying a ray-casting procedure where we simulate the human field of view (FOV). Such view frustum is obtained by using emitted rays starting from the estimated head position towards the corresponding estimated head orientation. The total illumination arriving to the person is computed by adding the related spatial illumination (radiance) from the patches of the scene that are in the direct visibility of the person. The rays project in the space as a uniform generated sequence over the unit sphere and weighted accordingly, based on the modelled luxmeter's LSC, towards the visible patches from the FOV of the sensor. The contribution of each patch to the total amount of lighting perceived by the occupant, is computed by estimating the percentage of rays intersecting that patch.

7.3 Invisible light switch evaluation

7.3.1 Dataset overview

[THCGD19], introduced a dataset for benchmarking light measurements with ground truth sensory data in real scenes. In this study we extended this dataset by introducing two more scenes with human activity, one based on a normal office environment and a second one representing a relaxing area (see Figure 7.3).

Both new scenes comprehend different human activities *e.g.* watching TV, working on a desk area, chatting, *etc.*, as well as different head orientations (VFOA) and multiple light combinations. In this work, VFOA is a cone with vertex in the middle of a person's eyes, oriented as the gaze direction and an aperture angle of $\alpha = 30°$.



(A) (B)

FIGURE 7.3: Illustration of the two indoor scenes used for evaluation: (a) illustrates a normal office environment and (b) shows a relaxing area. Red and green bounding boxes are showing the location of luxmeters within the space covering the spatial and gaze-gathered illumination ground truth measurements respectively.

In both rooms there is a controlled light management installation, where the position, type and properties (*e.g.* luminous intensity, light distribution curve, *etc.*) of the luminaires (eight in total) are considered known, see Figure 7.4.

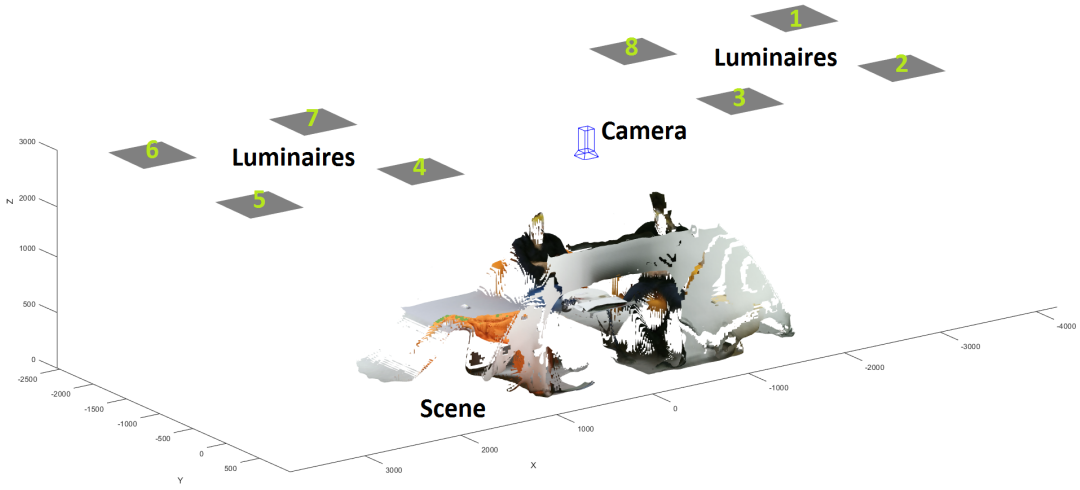


FIGURE 7.4: Illustration of the light management installation.

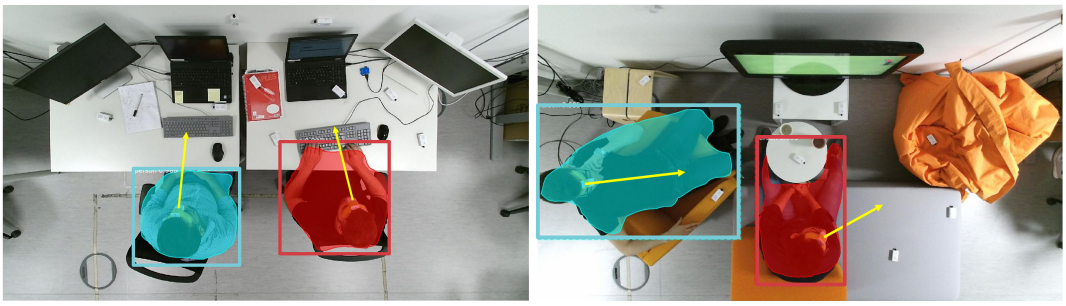
For obtaining the ground truth data we have installed and used a number of sensory equipment. A calibrated and aligned RGBD camera system (Kinect v2) is installed in the ceiling of the room providing a top-view perspective of the scene, see Fig. 7.3 and 7.4. Moreover, the camera is synchronized with a number of luxmeters (also indicated in Fig. 7.3) providing the light intensity ground truth data both for the spatial as well as for the gaze-gathered (attached to the forehead of the occupants) illumination. Considering the limitation (*i.e.* point-to-point) of lux readings that the luxmeters provide, we installed 11 sensors in different areas, thus providing a reasonable sampling of the scene. We use 9 luxemetes for evaluating the spatial illumination across the environment and 2 luxmeters for measuring the light intenisty that arrives to each one of the occupants appearing in the scenes. For each luxmeter, we additionally report the type and their specific light sensitivity characteristic curve, LSC (see Fig. 7.2) giving the sensor's sensitivity across the incident light angles.

Thereafter, we evaluate 24 and 30 different scenarios with different luminaire activations (luminaires switched on/off) for each room respectively (see Fig. 7.5). Our target was the use of RGB and depth input just for light measurement, the use of luxmeters as ground truth, and all other provided information for evaluation studies.

FIGURE 7.5: Illustration of 4 illumination variants within the two rooms. From left to right, the images illustrate the illumination provided by 1, 4, 7 and all 8 luminaires switched on in the two scenes.

7.3.2 Top-view detection and head-pose estimation

We fine tuned both the person detector and the head pose estimator on our top-view dataset. We report an average precision (AP) of 98% in terms of people detection. As mentioned previously we test our approach on the testing set of our top-view dataset. For the head pose orientation fine tuning on the whole body has been crucial for the performance, since using the sole head region produced definitely worst scores. In particular, we adopted two different class numbers for head pose, namely 4 and 8. The corresponding confusion matrices are reported in Fig. 7.7, showing an accuracy of 43.2% (8 classes) and 70.7% (4 classes) respectively. The scarce performance in the 8-class case was due to the mix among adjacent viewing angles: actually, the average size of the head region in the dataset is approx. 40x50 pixels. For these reasons, we used the 4-class version in the light perception studies.



(A) (B)

FIGURE 7.6: Illustration of people detection and head pose estimation. We detect people in the scene by using Mask R-CNN and then the detections are provided as input to the head pose estimator.

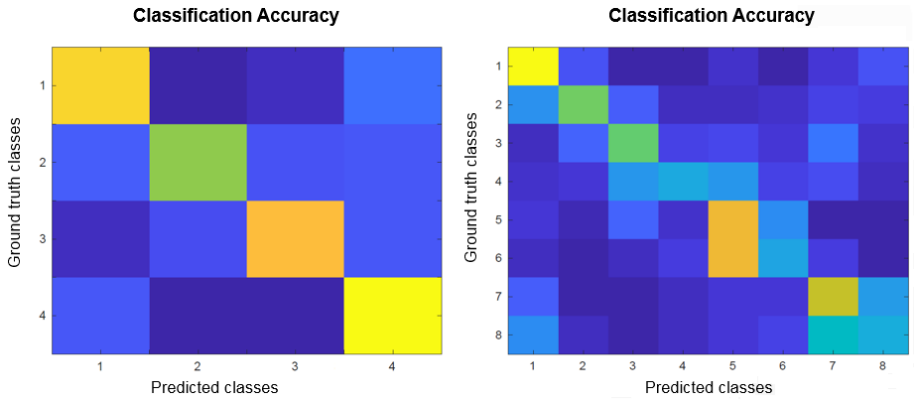


FIGURE 7.7: Confusion matrices of the head pose estimator. From left to right, the 4 and 8 classes confusion matrix respectively.

		Avg. error ε (in Lux)												
		Luxmeters												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Avg. (1-9)**	**Avg. (11-10)**
Scene 1	ε_{est} (w.r.t. GT)	62.5	26.3	68.0	65.1	47.9	57.1	44.0	29.9	28.0	97.6	92.2	56.2	94.7
	ε_{est_d} (w.r.t. GT)	-	-	-	-	-	-	-	-	-	216.08	166.4	-	191.24
Scene 2	ε_{est} (w.r.t. GT)	35.3	33.8	44.0	20.1	31.5	39.6	23.6	27.9	27.3	41.7	69.2	35.8	55.4
	ε_{est_d} (w.r.t. GT)	-	-	-	-	-	-	-	-	-	55.42	151.93	-	103.68

TABLE 7.1: The values represent the average estimated illumination error over the different lighting activation w.r.t. the ground truth measurements, for both scenes. Columns 1-9 corresponds to the spatial average values for the corresponding installed luxmeters in the environment. By contrast, values in columns 10-11 consider those luxmeters for evaluating the human light perception.

7.3.3 Person-perceived light estimation

Table 7.1 presents the quantitative results of our adopted light estimation approach. The table shows the average estimated error in lux values for both spatial (luxmeters 1-9) and gaze-gathered light estimation (luxmeters 10-11) cases. It can be easily noticed that the error, ε_{est}, for all luxmeters does not exceed the range of 100 lux, this yields an overall average light estimation error approx. 56 lux for Scene 1 and 36 lux for Scene 2. On the other hand, if we now consider only the luxmeters intended for evaluating the gaze-gathered light estimation, *i.e.* luxmeters 10 and 11, we notice that the error raises up to 94.7 lux and 55.4 lux for each scene respectively. This can be justified due to inaccuracies in the reconstruction of the 3D mesh areas corresponding to the head position and orientation of the occupants, as well as to the fact that the inter-reflections from the wall towards the sensors are limited due to incomplete reconstruction as an outcome of the limited FOV of the depth sensor. In any case, the fact that the average light estimation error does not exceed 100 lux indicates that the estimated illumination map can be considered reliable for describing the global illumination of the scene.

Furthermore, to demonstrate the applicability of our model, we used as explained a real person detector and a head pose estimator (making the pipeline completely automatic). In

Table 7.1 the ε_{est_d} rows for column 10 and 11, illustrates the error based on the detectors output for both scene 1 and 2. It can be observed that while the average error w.r.t. the oracle is less than 100 lux, this error raises up to the range of 200 lux negative variation w.r.t. to the ground truth measurements. The last can be justified by erroneous head pose estimations, considering the large step size (90°) of the 4-class adapted classification problem. This further brings into discussion the fact that this error could further be substantially reduced by improving the head pose estimator.

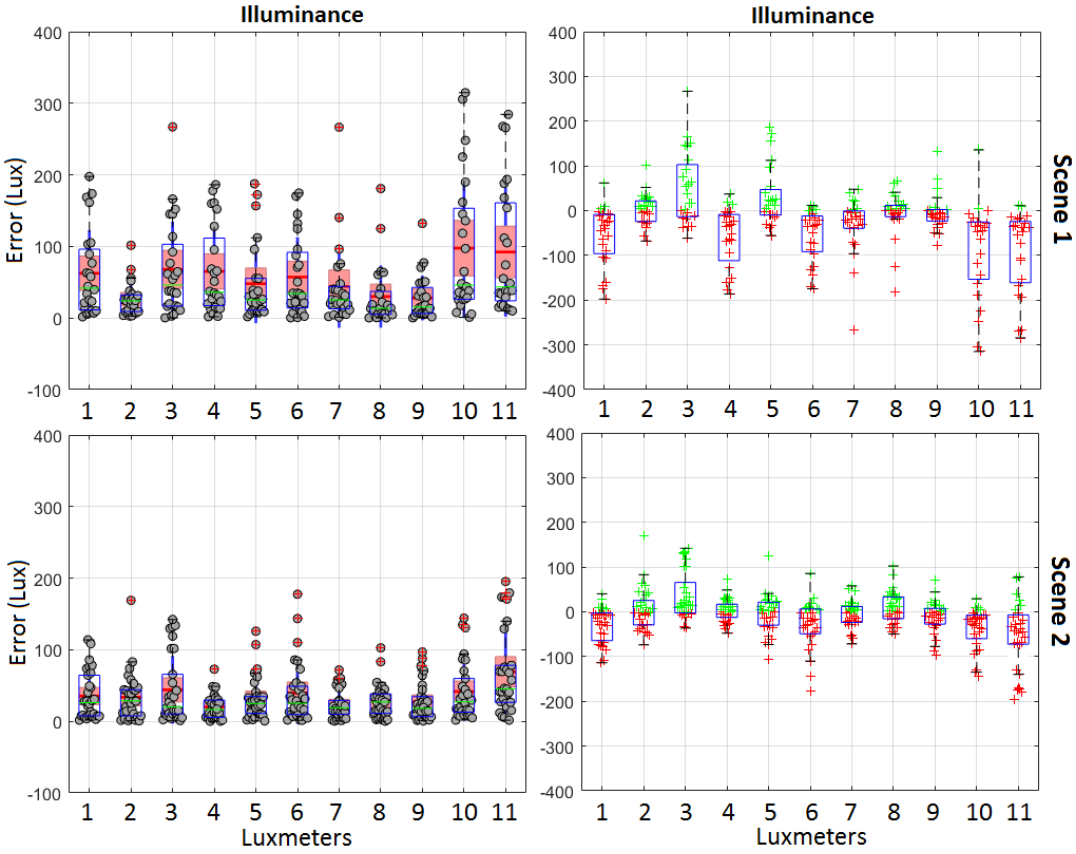


FIGURE 7.8: Scene 1 & 2 boxplot error evaluation (in Lux) using based on the presented framework. The boxplots in the first and second columns show the absolute and signed illumination estimated error for each lighting scenario in each scene respectively.

Figure 7.8 shows in a graph analysis the values presented in Table 7.1. The left graphs show the absolute light estimation error (y-axis), as estimated for each of the 11 (9 for spatial and 2 for the human light perception) used luxmeter sensors (x-axis). The gray dots, forming each of the box plot boxes, represent the estimated error of each of the lighting scenarios for each scene while the pink box represents the central 50% of the data. The upper and lower vertical lines indicate the extension of the remaining error points outside it and the central red line indicates the mean error which comes in alignment with the values shown in Table 7.1. Similarly, the boxplots on the right present the signed illumination error accordingly. The green and red markers indicate whether the error is due to an over or under estimation of the illuminance at the sensor's location respectively. As it can be noticed in the most of the cases the error is a result of an under estimation of the illuminance which as explained earlier are a cause of the incomplete geometry of the scenes as we only consider the parts of the environment within the FOV of the camera sensors.

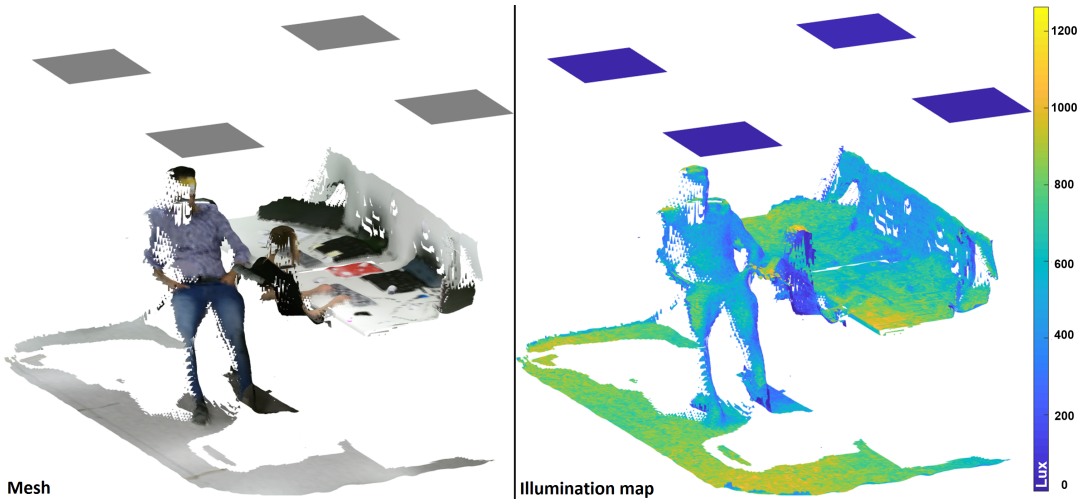


FIGURE 7.9: Illumination map of the full-lit scenario in scene 1 with a dense representation of the global illumination of the environment.

Finally, figures 7.9 and 7.10 visualise the illumination maps in the 3D space for one of the illumination scenarios in each of the scenes. As it can be seen the visualized illumination maps provide an accurate dense representation of the global illumination of the environment over time.

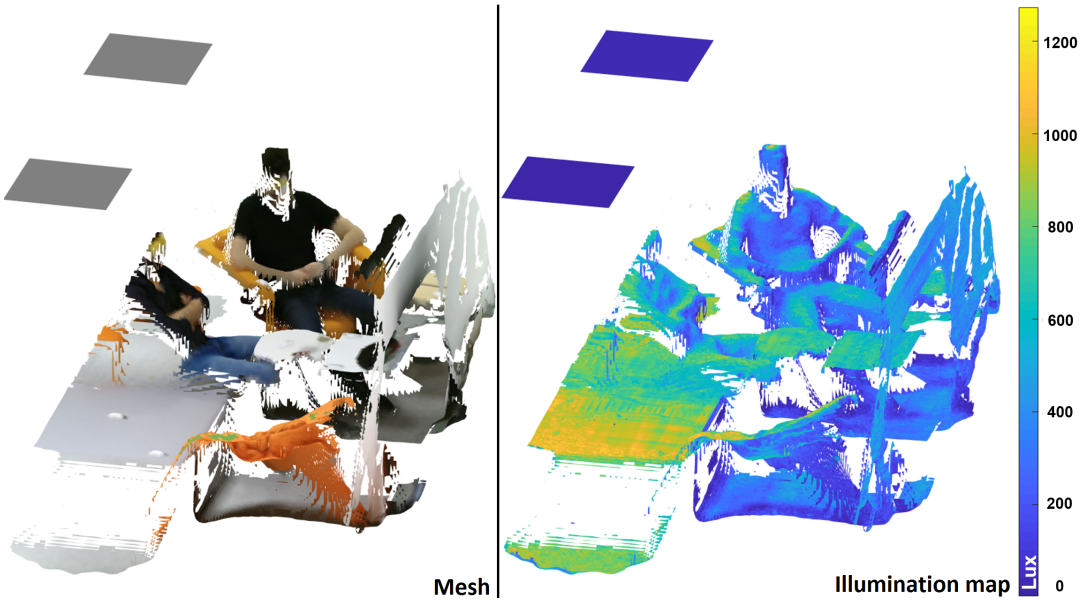


FIGURE 7.10: Illumination map of the full-lit scenario in scene 2. Notice the estimated illumination in the area in front of the occupants which is less bright in comparison to the one that are on their side. This is due to the body occlusion on the direct illumination coming from the luminaires from their back which is correctly estimated by the ILS.

		Scene 1						Scene 2						
		VFOA 1			VFOA 2			VFOA 1			VFOA 2			
Luminaire activations		**3|4|7|8**	**2|3|4|5**	**3|4**	**3|4|7|8**	**2|3|4|5**	**3|4**	**3|4|7|8**	**2|3|4|5**	**3|4**	**1|2|3|4|5|6**	**2|3|4|5**	**1|3|4|6**	**3|4**
Luxmeter 10	Δ_{lux} **(w.r.t. full-lit)**	116.15	123.77	189.01	85.4	123.8	163.85	84.23	93.69	151.92	106.52	148.12	157.07	191.15
	ε_{est} **(w.r.t. GT)**	167.2	144.09	102.73	235.3	200.1	163.28	85.85	94.1	43.76	22.94	12.97	13.59	25.69
	Δ_{watt} **(w.r.t. full-lit)**	387.2	387.2	580.8	387.2	387.2	580.8	387.2	387.2	580.8	193.6	387.2	387.2	580.8
Luxmeter 11	Δ_{lux} **(w.r.t. full-lit)**	97.68	125.15	169.72	167.4	86.34	194.37	62.67	118.21	153.02	99.17	154.28	167.93	194.85
	ε_{est} **(w.r.t. GT)**	194.63	171.74	131.55	91.14	128.7	70.21	15.26	67.87	5.39	9.4	241.12	2.81	203.69
	Δ_{watt} **(w.r.t. full-lit)**	387.2	387.2	580.8	387.2	387.2	580.8	387.2	387.2	580.8	193.6	387.2	387.2	580.8

TABLE 7.2: Quantitative analysis of four different head orientation class studies (VFOA), two for each scene. Δ_{lux} shows the discrepancy of different lighting scenarios w.r.t. the full lit scenario (reference). ε_{est} shows the corresponding average error of the estimated light in regards to the ground truth lux measurements and Δ_{watt} shows the discrepancy of the power consumption in watts considering the active/non active luminaires for each corresponding scenario.

7.3.4 Applications of the invisible light switch

The idea of the Invisible Light Switch is straightforward as we have presented above. Thus, in Table 7.2 we examine the applicability of the invisible light switch from the human perspective aspect (luxmeters 10-11) for different head orientation cases (VFOA) in the two scenes. The value Δ_{lux} provides the information regarding what is the impact to the light perceived from the occupants (based on the ground truth sensor measurements) on different light source combination scenarios. As it can be seen this gives us a range of 0-200 lux negative variation even to the most aggressive scenario of having only two luminaires active (the ones to the direct view of the occupants each time). If we connect this with the amount of watts that we can save for this corresponding lighting scenario, *i.e.* $\Delta_{watt} = 580.8$ watt w.r.t. to the full lit case, this can give us a total power efficiency of 12379.2 KWatt through a whole day. The value ε_{est} reports the light estimation error based on our framework, which as we can see again it settles within a range of 0-200 lux overall negative variation. This error shows us how our system aligns with the ground truth measurements, *i.e.* a lower ε_{est} error the better, and whether the same pattern described above could be followed. A visual example of the VFOA 1 case for scene 1 (see Table 7.2) can be seen in Figure 7.11. As it can be easily noticed the estimated illumination over the desk areas have the less affect as we switch off the peripheral light sources and still providing an optimally lit scenario while it is minimally lit.

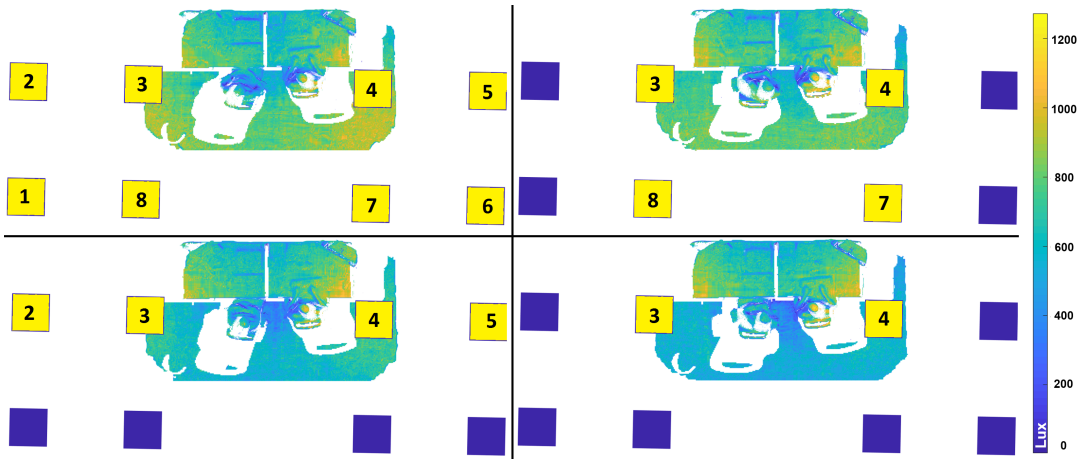


FIGURE 7.11: Qualitative illustration of the VFOA 1 ablation study for Scene 1 presented in table 7.2. The top left corner shows the illumination map of the full lit case, in comparison to three other light scenarios. As it can be seen the estimated illumination over the desk area where the two occupants have their attention is less affected in comparison to the areas behind them. This show in practice how the invisible light switch application could be established.

Chapter 8

Summary and Conclusion

The main aim of this thesis, is to justify the effort in the estimation of the head pose estimation in surveillance scenarios. Other than being useful for individuating groups (or interactive activities) or highlighting salient areas in the scene, the head pose serves to individuate the visual frustum of attention, which in turns is shown to be indicative for guessing the future path of pedestrian.

Scenarios such as surveillance, where precise gazing information cannot be retrieved head pose serves as a proxy to visual frustum of attention. To this end, we proposed a CNN pipeline that copes simultaneously with pedestrian detection and head pose estimation, in surveillance scenarios. We demonstrated that the joint model performs competitively with the state-of-the-art, beating up-to-date serial pipelines composed by pedestrian detectors, head detectors and head pose estimators. At the same time, we confirmed that the body information is an important cue to increase performance of head pose estimation, especially when the head patch size is small.

Furthermore, We have argued for the importance of people head poses, as encoded in the proposed vislets, to forecast their future motion.We have shown that vislets are mostly aligned with the people motion, and therefore useful to forecast it. But when vislets are not aligned with the people motion, then they express the intention of people to change direction. Vislets differ from the current approaches, as most recent LSTM-based forecasting has only considered own and neighboring pedestrian positions. But this is close in spirit to decade-old works using the people desired goals. In this work, the head pose is however estimated, not provided (e.g. by an oracle). The use of vislets is enabled by the novel MX-LSTM framework. This jointly "reasons" on tracklets and vislets by means of a multi-variate Gaussian distribution, the covariance of which encodes the interplay of position and head pose. Our proposed log-Cholesky parameterization allows its unconstrained optimization by the LSTM backpropagation, and it opens the way to including additional variables (e.g. the people belonging to a social group). Finally, this work has delved into a comprehensive evaluation of the proposed MX-LSTM, including ablation studies on vislets (both estimated and provided as GT), social pooling, view frustum, observation and prediction time horizons. MX-LSTM provides currently state of the art performance and it is most effective when people slow down and look around to change direction, the Achilles heel of other modern techniques.

Finally, in this thesis we have proposed a practical application for smart lighting. We have proposed an Invisible Light Switch. The idea behind the Invisible Light Switch is straightforward: the user controls and sets the illumination of the environment that he can see (estimated by VFOA), while the proposed system acts on the part of the environment that the user cannot see, turning off the lights, thus ensuring a consistent energy saving. The study of the scene as discussed above serves this goal: knowing the 3D geometry of the scene and the map of inter-reflectance will allow to understand how the different light sources impact each point of the space; knowing where a user is located and what is his posture serves to infer what he can see and what he cannot, individuating potential areas where the light can

be turned off. Being able to forecast his future activities will help understand (in advance) which lights should be turned on, avoiding the user to continuously act on the illumination system, and showing the user the illumination scenario that he wants to have.

Bibliography

[ABW06] Gianluca Antonini, Michel Bierlaire, and Mats Weber. “Discrete choice models of pedestrian walking behavior”. In: *Transportation Research Part B: Methodological* 40.8 (2006), pp. 667–687.

[AFDP13] Ron Appel et al. “Quickly boosting decision trees–pruning underachieving features early”. In: *International conference on machine learning*. 2013, pp. 594–602.

[AGRRF+16] A. Alahi et al. “Social LSTM: Human Trajectory Prediction in Crowded Spaces”. In: *CVPR*. 2016.

[Aka69] H. Akaike. “Fitting autoregressive models for prediction”. In: *Annals of the institute of Statistical Mathematics* 21.1 (1969), pp. 243–247.

[AMBT06] G. Antonini et al. “Behavioral priors for detection and tracking of pedestrians in video sequences”. In: *International Journal of Computer Vision* 69 (2006).

[AN04] P. Abbeel and A. Y. Ng. “Apprenticeship learning via inverse reinforcement learning”. In: *ICML*. 2004.

[ARF14] A. Alahi, V. Ramanathan, and L. Fei-Fei. “Socially-aware large-scale crowd forecasting”. In: *CVPR*. 2014.

[BCTFP+13] L. Bazzani et al. “Social Interactions by Visual Focus of Attention in a Three-Dimensional Environment”. In: *Expert Systems* 30.2 (2013), pp. 115–127.

[BFS17] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. “Long-Term On-Board Prediction of Pedestrians in Traffic Scenes”. In: *1st Conference on Robot Learning*. 2017.

[BHHA18] Stefan Becker et al. “An Evaluation of Trajectory Prediction Approaches and Notes on the TrajNet Benchmark”. In: *arXiv preprint arXiv:1805.07663* (2018).

[BO04] S. O. Ba and J.-M. Odobez. “A Probabilistic Framework for Joint Head Tracking and Pose Estimation”. In: *ICPR*. 2004.

[BR09a] B. Benfold and I. Reid. “Guiding Visual Surveillance by Tracking Human Attention”. In: *BMVC*. 2009.

[BR09b] Ben Benfold and Ian Reid. “Guiding Visual Surveillance by Tracking Human Attention”. In: *Proceedings of the 20th British Machine Vision Conference*. Sept. 2009.

[BR11] Ben Benfold and Ian Reid. “Stable multi-target tracking in real-time surveillance video”. In: *CVPR*. 2011.

[BX05] Stephen Boyd and Lin Xiao. “Least-squares covariance matrix adjustment”. In: *SIAM Journal on Matrix Analysis and Applications* 27.2 (2005), pp. 532–546.

[CADNT17] Huseyin Coskun et al. “Long Short-Term Memory Kalman Filters: Recurrent Neural Estimators for Pose Regularization”. In: *ICCV*. 2017.

[CBPFT+11a] Marco Cristani et al. “Social interaction discovery by statistical analysis of F-formations.” In: *BMVC*. Vol. 2. 2011, p. 4.

[CBPFT+11b] M. Cristani et al. “Social interaction discovery by statistical analysis of F-formations”. In: *BMVC*. 2011.

[CLFK01] Robert T Collins et al. “Algorithms for cooperative multisensor surveillance”. In: *Proceedings of the IEEE* 89.10 (2001), pp. 1456–1477.

[CO12] Cheng Chen and Jean-Marc Odobez. “We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video”. In: *CVPR*. 2012.

[CPKMY18] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018), pp. 834–848.

[CRSBC+15] Davide Conigliaro et al. “The s-hock dataset: Analyzing crowds at the stadium”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2039–2047.

[CS12] Wongun Choi and Silvio Savarese. “A unified framework for multi-target tracking and collective activity recognition”. In: *European Conference on Computer Vision*. Springer. 2012, pp. 215–230.

[CS14] Wongun Choi and Silvio Savarese. “Understanding collective activities of people from videos”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.6 (2014), pp. 1242–1257.

[CU90] U. Castiello and C. Umiltà. “Size of the Attentional Focus and Efficiency of Processing”. In: *Acta psychologica* 73.3 (1990), pp. 195–209.

[Cv80] J. F. Caminada and W. J. M. van Bommel. *Philips Engineering Report 43*. 1980.

[DABP14] Piotr Dollar et al. “Fast Feature Pyramids for Object Detection”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.8 (Aug. 2014), pp. 1532–1545. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2300479. URL: http://dx.doi.org/10.1109/TPAMI.2014.2300479.

[Dag11] Gislin Dagnelie. *Visual Prosthetics: Physiology, Bioengineering, Rehabilitation*. Springer Science & Business Media, 2011.

[DLB10] C. Djeraba, A. Lablack, and Y. Benabbas. *Multi-Modal User Interactions in Controlled Environments*. Multimedia Systems and Applications. Springer US, 2010. ISBN: 9781441903167. URL: https://books.google.it/books?id=PMC0hzuYvYYC.

[DR12] N. Davoudian and P. Raynham. “What do Pedestrians Look at at Night?” In: *Lighting Research and Technology* 44.4 (2012), pp. 438–448.

[DRS11] A. D. Dragan, N. D. Ratliff, and S. S. Srinivasa. “Manipulation planning with goal sets using constrained trajectory optimization”. In: *ICRA*. 2011.

[DS96] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[DWSP09] Piotr Dollár et al. “Pedestrian detection: A benchmark”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 304–311.

[DWW15] Yong Du, Wei Wang, and Liang Wang. “Hierarchical recurrent neural network for skeleton based action recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1110–1118.

[Eng94] Basil G Englis. “The role of affect in political advertising: Voter emotional responses to the nonverbal behavior of politicians”. In: *Attention, attitude, and affect in response to advertising* (1994), pp. 223–247.

[EVWWZ10] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[EZWV06] Mark Everingham et al. “The pascal visual object classes challenge 2006 (voc 2006) results”. In: (2006).

[FFK13] T. Foulsham, J. Farley, and A. Kingstone. “Mind wandering in sentence reading: Decoupling the link between mind and eye”. In: *Canadian Journal of Experimental Psychology* 67.1 (2013), p. 51.

[FG03] J. M. Findlay and I. D. Gilchrist. *Active vision: The psychology of looking and seeing*. Oxford University Press, 2003.

[FGM10] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. “Cascade object detection with deformable part models”. In: *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE. 2010, pp. 2241–2248.

[FGMR10] Pedro F Felzenszwalb et al. “Object detection with discriminatively trained part-based models”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2010), pp. 1627–1645.

[FHSM79] John E Flynn et al. “A guide to methodology procedures for measuring subjective impressions in lighting”. In: *Journal of the Illuminating Engineering Society* 8.2 (1979), pp. 95–110.

[FHT+00] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)”. In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[FMFGL+96] David A Forsyth et al. “Finding pictures of objects in large collections of images”. In: *International workshop on object representation in computer vision*. Springer. 1996, pp. 335–360.

[FMW00] James M Ferryman, Stephen J Maybank, and Anthony D Worrall. “Visual surveillance for moving vehicles”. In: *International Journal of Computer Vision* 37.2 (2000), pp. 187–197.

[FUCH15] S. Fotios et al. “Using eye-tracking to identify pedestrians’ critical visual tasks, Part 1. Dual task approach”. In: *Lighting Research & Technology* 47.2 (2015), pp. 133–148.

[FUY15] S. Fotios, J. Uttley, and B. Yang. “Using eye-tracking to identify pedestrians’ critical visual tasks. Part 2. Fixation on pedestrians”. In: *Lighting Research & Technology* 47.2 (2015), pp. 149–160.

[FWK11] T. Foulsham, E. Walker, and A. Kingstone. “The where, what and when of gaze allocation in the lab and the natural environment”. In: *Vision research* 51.17 (2011), pp. 1920–1931.

[GDDM14] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[GDGRW15] Karol Gregor et al. “DRAW: A recurrent neural network for image generation”. In: *arXiv preprint arXiv:1502.04623* (2015).

[Gir15] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[GJFSA18a] Agrim Gupta et al. “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. CONF. 2018.

[GJFSA18b] Agrim Gupta et al. “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. CONF. 2018.

[GMHC06] Nicolas Gourier et al. “Head pose estimation on low resolution images”. In: *CLEAR*. 2006.

[Gra12] Alex Graves. *Supervised sequence labelling with recurrent neural networks*. Vol. 385. Springer, 2012.

[Gra13] Alex Graves. “Generating sequences with recurrent neural networks”. In: *arXiv preprint arXiv:1308.0850* (2013).

[GS05] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural Networks* 18.5-6 (2005), pp. 602–610.

[GSKSS17] Klaus Greff et al. “LSTM: A search space odyssey”. In: *IEEE transactions on neural networks and learning systems* 28.10 (2017), pp. 2222–2232.

[GV06] Anca D Galasiu and Jennifer A Veitch. “Occupant preferences and satisfaction with the luminous environment and control systems in daylit offices: a literature review”. In: *Energy and Buildings* 38.7 (2006), pp. 728–742.

[GXH10] Shaogang Gong, Tao Xiang, and Somboon Hongeng. “Learning Human Pose in Crowd”. In: *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*. MPVA ’10. Firenze, Italy: ACM, 2010, pp. 47–52. ISBN: 978-1-4503-0167-1. DOI: 10.1145/1878039.1878050. URL: http://doi.acm.org/10.1145/1878039.1878050.

[Hal66] Edward Twitchell Hall. *The hidden dimension*. Doubleday & Co, 1966.

[HBFS+01] Sepp Hochreiter et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

[HD14] Minh Hoai and Fernando De la Torre. “Max-margin early event detectors”. In: *International Journal of Computer Vision* 107.2 (2014), pp. 191–202.

[Hig88] Nicholas J Higham. “Computing a nearest symmetric positive semidefinite matrix”. In: *Linear algebra and its applications* 103 (1988), pp. 103–118.

[HLZHW+16] Siyu Huang et al. “Deep learning driven visual path prediction from a single image”. In: *IEEE Transactions on Image Processing* 25.12 (2016), pp. 5892–5904.

[HM95] D. Helbing and P. Molnar. “Social force model for”. In: *Physical review E* 51.5 (1995), p. 4282.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[HSTDC+18] Irtiza Hasan et al. ““Seeing is Believing”: Pedestrian Trajectory Forecasting Using Visual Frustum of Attention”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018.

[HSTDG+18] Irtiza Hasan et al. “MX-LSTM: mixing tracklets and vislets to jointly forecast trajectories and head poses”. In: 2018.

[HTGDC17] I. Hasan et al. “Tiny head pose classification by bodily cues”. In: *ICIP*. 2017.

[IC01] J. Intriligator and P. Cavanagh. “The spatial resolution of visual attention”. In: *Cognitive psychology* 43.3 (2001), pp. 171–216.

[JL16] Huaizu Jiang and Erik G. Learned-Miller. “Face Detection with the Faster R-CNN”. In: *CoRR* abs/1606.03473 (2016). URL: http://arxiv.org/abs/1606.03473.

[JS01] S Rao Jammalamadaka and Ambar Sengupta. *Topics in circular statistics*. Vol. 5. World Scientific, 2001.

[KA15] D Kinga and J Ba Adam. “A method for stochastic optimization”. In: *International Conference on Learning Representations (ICLR)*. Vol. 5. 2015.

[Kal+60] Rudolph Emil Kalman et al. “A new approach to linear filtering and prediction problems”. In: *ASME Journal of Basic Engineering* (1960).

[Kaw08] Kazuya Kawakami. “Supervised Sequence Labelling with Recurrent Neural Networks”. PhD thesis. PhD thesis. Ph. D. thesis, Technical University of Munich, 2008.

[Ken67] A. Kendon. “Some functions of gaze-direction in social interaction”. In: *Acta psychologica* 26 (1967), pp. 22–63.

[Ken90] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. Cambridge University Press, 1990.

[KKSB12] Markus Kuderer et al. “Feature-Based Prediction of Trajectories for Socially Compliant Navigation.” In: *Robotics: science and systems*. 2012.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[KZBH12] K. Kitani et al. “Activity forecasting”. In: *ECCV*. 2012.

[LCL07] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. “Crowds by example”. In: *Computer Graphics Forum*. 2007.

[LCVCT+17] Namhoon Lee et al. “Desire: Distant future prediction in dynamic scenes with interacting agents”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 336–345.

[LJMMH17] Stéphane Lathuilière et al. “Deep mixture of linear inverse regressions applied to head-pose estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4817–4825.

[LK16] N. Lee and K. M. Kitani. “Predicting wide receiver trajectories in american football”. In: *WACV*. 2016.

[LLSXF+15] Jianan Li et al. “Scale-aware fast R-CNN for pedestrian detection”. In: *arXiv preprint arXiv:1510.08160* (2015).

[LLWLP16] Xiabing Liu et al. “3D head pose estimation with convolutional neural network trained on synthetic images”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 1289–1293.

[LMBHP+14] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[LOWFC+18] Li Liu et al. “Deep learning for generic object detection: A survey”. In: *arXiv preprint arXiv:1809.02165* (2018).

[LPR11] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. “Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker”. In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE. 2011, pp. 120–127.

[LRWW98] J. C. Lagarias et al. “Convergence properties of the Nelder–Mead simplex method in low dimensions”. In: *SIAM Journal on Optimization* 9.1 (1998), pp. 112–147.

[LSXW16] Jun Liu et al. “Spatio-temporal lstm with trust gates for 3d human action recognition”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 816–833.

[LYO15a] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. “Fast and accurate head pose estimation via random projection forests”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1958–1966.

[LYO15b] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. “Fast and accurate head pose estimation via random projection forests”. In: *ICCV*. 2015.

[LYU18] Wenjie Luo, Bin Yang, and Raquel Urtasun. “Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting With a Single Convolutional Net”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3569–3577.

[MHB16] J. Mainprice, R. Hayne, and D. Berenson. “Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces”. In: *IEEE Trans. on Robotics* 32.4 (2016), pp. 897–908.

[MHLK17] Wei-Chiu Ma et al. “Forecasting interactive dynamics of pedestrians with fictitious play”. In: *CVPR*. 2017.

[MN89] P. McCullagh and J. A. Nelder. *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*. 1989.

[MP07] D. S. Marigold and A. E. Patla. “Gaze fixation patterns for negotiating complex ground terrain”. In: *Neuroscience* 144.1 (2007), pp. 302–313.

[MR15] Sankha S Mukherjee and Neil Martin Robertson. “Deep head pose: Gaze-direction estimation in multimodal video”. In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2094–2107.

[MT08] Brendan Tran Morris and Mohan Manubhai Trivedi. “A survey of vision-based trajectory learning and analysis for surveillance”. In: *IEEE Trans. on Circuits and Systems for Video Technology* 18.8 (2008), pp. 1114–1127.

[MT09] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. “Head pose estimation in computer vision: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.4 (2009), pp. 607–626.

[MWF16] Andrii Maksai, Xinchao Wang, and Pascal Fua. "What players do with the ball: A physically constrained interaction modeling". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 972–981.

[MWFF17] Andrii Maksai et al. "Non-Markovian Globally Consistent Multi-Object Tracking". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2017, pp. 2563–2573.

[NDH14] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. "Local decorrelation for improved pedestrian detection". In: *Advances in Neural Information Processing Systems*. 2014, pp. 424–432.

[OCM07] Margarita Osadchy, Yann Le Cun, and Matthew L Miller. "Synergistic face detection and pose estimation with energy-based models". In: *Journal of Machine Learning Research* 8.May (2007), pp. 1197–1215.

[OGX09] Javier Orozco, Shaogang Gong, and Tao Xiang. "Head pose classification in crowded scenes." In: *BMVC*. Vol. 1. 2009, p. 3.

[PB96] José C Pinheiro and Douglas M Bates. "Unconstrained parametrizations for variance-covariance matrices". In: *Statistics and Computing* 6.3 (1996), pp. 289–296.

[PESV09] S. Pellegrini et al. "You'll never walk alone: Modeling social behavior for multi-target tracking". In: *ICCV*. 2009.

[Pou11] Mohsen Pourahmadi. "Covariance estimation: The GLM and regularization perspectives". In: *Statistical Science* (2011), pp. 369–387.

[Pri81] M. B. Priestley. *Spectral analysis and time series*. Academic press, 1981.

[PV03] A. E. Patla and J. N. Vickers. "How far ahead do we look when required to step on specific locations in the travel path during locomotion?" In: *Experimental brain research* 148.1 (2003), pp. 133–138.

[QR05] J. Quiñonero-Candela and C. E. Rasmussen. "A unifying view of sparse approximate Gaussian process regression". In: *Journal of Machine Learning Research* 6.12 (2005), pp. 1939–1959.

[Ras06] C. E. Rasmussen. "Gaussian processes for machine learning". In: *Adaptive Computation and Machine Learning*. 2006.

[RDSKS+15] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[RF87] AJ Robinson and Frank Fallside. *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering, 1987.

[RHGS15a] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *CoRR* abs/1506.01497 (2015). URL: http://arxiv.org/abs/1506.01497.

[RHGS15b] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[RHGS17] Shaoqing Ren et al. "Faster R-CNN: towards real-time object detection with region proposal networks". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (2017), pp. 1137–1149.

[RPC19] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1 (2019), pp. 121–135.

[RR06a] N. M. Robertson and I. D. Reid. “Estimating Gaze Direction from Low-Resolution Faces in Video”. In: *ECCV*. 2006.

[RR06b] N. Robertson and I. Reid. “Estimating Gaze Direction from Low-Resolution Faces in Video”. In: *European Conference on Computer Vision (ECCV)*. 2006.

[RR11] N. M. Robertson and I. D. Reid. “Automatic Reasoning about Causal Events in Surveillance Video”. In: *EURASIP Journal on Image and Video Processing* 2011.1 (2011), p. 530325.

[RSAS16] A. Robicquet et al. “Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes”. In: *ECCV*. 2016.

[RSRVL+14] Anoop Kolar Rajagopal et al. “Exploring transfer learning approaches for head pose classification from multi-view surveillance images”. In: *International journal of computer vision* 109.1-2 (2014), pp. 146–167.

[Ryo11] Michael S Ryoo. “Human activity prediction: Early recognition of ongoing activities from streaming videos”. In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 1036–1043.

[SAS17] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. “Tracking the untrackable: Learning to track multiple cues with long-term dependencies”. In: *arXiv preprint arXiv:1701.01909* (2017).

[SBOG08] K. Smith et al. “Tracking the visual focus of attention for a varying number of wandering people”. In: *IEEE TPAMI* 30.7 (2008), pp. 1212–1229.

[SDZLZ16] Hang Su et al. “Crowd Scene Understanding with Coherent Recurrent Neural Networks”. In: *IJCAI*. 2016.

[SFYW99] R. Stiefelhagen et al. “From gaze to focus of attention”. In: *VISUAL*. 1999.

[SMS15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. “Unsupervised learning of video representations using lstms”. In: *International conference on machine learning*. 2015, pp. 843–852.

[SYMHD17] Li Sun et al. “3DOF Pedestrian Trajectory Prediction Learned from Long-Term Autonomous Mobile Robot Deployment Data”. In: *arXiv preprint arXiv:1710.00126* (2017).

[SZDZ17] Hang Su et al. “Forecast the plausible paths in crowd scenes”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 2772–2778.

[TCP06] Adrien Treuille, Seth Cooper, and Zoran Popović. “Continuum crowds”. In: *ACM Transactions on Graphics (TOG)* 25.3 (2006), pp. 1160–1168.

[TFSMC10] Diego Tosato et al. “Multi-class classification on riemannian manifolds for video surveillance”. In: *European conference on computer vision*. Springer. 2010, pp. 378–391.

[THCGD19] Theodore Tsesmelis et al. “RGBD2lux: Dense light intensity estimation with an RGBD sensor”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 501–510.

[TK10] P. Trautman and A. Krause. “Unfreezing the robot: Navigation in dense, interacting crowds”. In: *IROS*. 2010.

[TLWT15] Yonglong Tian et al. “Deep learning strong parts for pedestrian detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1904–1912.

[TSCM13a] Diego Tosato et al. “Characterizing Humans on Riemannian Manifolds”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (Aug. 2013), pp. 1972–1984. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.263. URL: http://dx.doi.org/10.1109/TPAMI.2012.263.

[TSCM13b] D. Tosato et al. “Characterizing Humans on Riemannian Manifolds”. In: *IEEE TPAMI* 35.8 (2013), pp. 1972–1984.

[UVGS13] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[VCDPL13] P. Vansteenkiste et al. “The visual control of bicycle steering: The effects of speed and path width”. In: *Accident Analysis & Prevention* 51 (2013), pp. 222–227.

[VMCHP+16] S. Vascon et al. “Detecting conversational groups in images and sequences: A robust game-theoretic approach”. In: *Computer Vision and Image Understanding* 143 (2016), pp. 11–24.

[VS17] Daksh Varshneya and G. Srinivasaraghavan. “Human Trajectory Prediction using Spatially aware Deep Attention Models”. In: *NIPS*. 2017.

[VTBE15] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *CVPR*. 2015.

[WFH08] J. M. Wang, D. J. Fleet, and A. Hertzmann. “Gaussian process dynamical models for human motion”. In: *IEEE TPAMI* 30.2 (2008), pp. 283–298.

[Wil89] Ronald J Williams. *Complexity of exact gradient computation algorithms for recurrent neural networks*. Tech. rep. Technical Report Technical Report NU-CCS-89-27, Boston: Northeastern . . ., 1989.

[Wil98] C. K. I. Williams. “Prediction with Gaussian processes: From linear regression to linear prediction and beyond”. In: *Learning in graphical models*. Springer, 1998, pp. 599–621.

[XHR18] Hao Xue, Du Q Huynh, and Mark Reynolds. “SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 1186–1194.

[XPDY08] Zhengzheng Xing et al. “Mining sequence classifiers for early prediction”. In: *Proceedings of the 2008 SIAM international conference on data mining*. SIAM. 2008, pp. 644–655.

[YBOB11] K. Yamaguchi et al. “Who are you with and where are you going?” In: *CVPR*. 2011.

[YLW15] S. Yi, H. Li, and X. Wang. “Understanding pedestrian behaviors from stationary crowd groups”. In: *CVPR*. 2015.

[YYJ15] CAI Ying, Meng-long Yang, and LI Jun. “Multiclass classification based on a deep convolutional”. In: *Frontiers of Information Technology & Electronic Engineering* 16.11 (2015), pp. 930–939.

[YYLL14a] Bin Yang et al. “Aggregate channel features for multi-view face detection”. In: *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE. 2014, pp. 1–8.

[YYLL14b] Bin Yang et al. “Aggregate channel features for multi-view face detection”. In: *CoRR* abs/1407.4023 (2014). URL: http://arxiv.org/abs/1407.4023.

[ZBS15] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. “Filtered channel features for pedestrian detection”. In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE. 2015, pp. 1751–1760.

[ZBS17] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. “Citypersons: A diverse dataset for pedestrian detection”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2. 2017, p. 3.

[ZH16] L. Zhang and H. Hung. “Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images”. In: *CVPR*. 2016.

[ZKLOT14] Bolei Zhou et al. “Object detectors emerge in deep scene cnns”. In: *arXiv preprint arXiv:1412.6856* (2014).

[ZKLOT16] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.

[ZLLH16] Liliang Zhang et al. “Is Faster R-CNN Doing Well for Pedestrian Detection?” In: *European Conference on Computer Vision*. Springer. 2016, pp. 443–457.

[ZMBD08] Brian D. Ziebart et al. “Maximum Entropy Inverse Reinforcement Learning”. In: *AAAI*. 2008.

[ZRGMP+09] Brian D Ziebart et al. “Planning-based prediction for pedestrians”. In: *IROS*. 2009.