

# THÈSE DE DOCTORAT

Suivi long terme de personnes pour les  
systèmes de vidéo monitoring

Long-term people trackers for video monitoring systems

**Thi Lan Anh NGUYEN**

INRIA Sophia Antipolis, France

**Présentée en vue de l'obtention  
du grade de docteur en Informatiques  
d'Université Côte d'Azur**  
**Dirigée par** : Francois Bremond  
**Soutenu le** : 17/07/2018

**Devant le jury, composé de :**

- Frederic Precioso, Professor, I3S lab – France
- Francois Bremond, Team leader, INRIA Sophia Antipolis – France
- Jean-Marc Odobez, Team leader, IDIAP – Switzerland
- Jordi Gonzalez, Associate Professor, ISE lab, Espanol
- Serge Miguet, Professor, ICOM, Université Lumière Lyon 2, France



# Suivi long terme de personnes pour les systèmes de vidéo monitoring

Long-term people trackers for video monitoring systems

Jury:

Président du jury\*

Frederic Prescioso, Professor, I3S lab - France

Rapporteurs

Jean-Mard Odobez, Team leader, IDIAP – Swizerland

Jordi Gonzales, Associate Professor, ISE lab, Espagnol

Serge Miguet, Professor, ICOM, Universite Lumiere Lyon 2 – France

Directeur de thèse :

Francois Bremond, Team leader, STARS team, INRIA Sophia Antipolis

## **Titre : Suivi long terme de personnes pour les systèmes de vidéo monitoring**

---

### **Résumé**

Le suivi d'objets multiples (Multiple Object Tracking (MOT)) est une tâche importante dans le domaine de la vision par ordinateur. Plusieurs facteurs tels que les occlusions, l'éclairage et les densités d'objets restent des problèmes ouverts pour le MOT. Par conséquent, cette thèse propose trois approches MOT qui se distinguent à travers deux propriétés: leur généralité et leur efficacité.

La première approche sélectionne automatiquement les primitives visuelles les plus fiables pour caractériser chaque tracklet dans une scène vidéo. Aucun processus d'apprentissage n'est nécessaire, ce qui rend cet algorithme générique et déployable pour une grande variété de systèmes de suivi.

La seconde méthode règle les paramètres de suivi en ligne pour chaque tracklet, en fonction de la variation du contexte qui l'entoure. Il n'y a pas de contraintes sur le nombre de paramètres de suivi et sur leur dépendance mutuelle. Cependant, on a besoin de données d'apprentissage suffisamment représentatives pour rendre cet algorithme générique.

La troisième approche tire pleinement avantage des primitives visuelles (définies manuellement ou apprises), et des métriques définies sur les tracklets, proposées pour la ré-identification et leur adaptation au MOT. L'approche peut fonctionner avec ou sans étape d'apprentissage en fonction de la métrique utilisée.

Les expériences sur trois ensembles de vidéos, MOT2015, MOT2017 et ParkingLot montrent que la troisième approche est la plus efficace. L'algorithme MOT le plus approprié peut être sélectionné, en fonction de l'application choisie et de la disponibilité de l'ensemble des données d'apprentissage.

---

**Mots clés :** MOT, suivi de personnes

---

**Title:** Long term people trackers for video monitoring systems

---

### **Abstract**

Multiple Object Tracking (MOT) is an important computer vision task and many MOT issues are still unsolved. Factors such as occlusions, illumination, object densities are big challenges for MOT. Therefore, this thesis proposes three MOT approaches to handle these challenges. The proposed approaches can be distinguished through two properties: their generality and their effectiveness.

The first approach selects automatically the most reliable features to characterize each tracklet in a video scene. No training process is needed which makes this algorithm generic and deployable within a large variety of tracking frameworks. The second method tunes online tracking parameters for each tracklet according to the variation of the tracklet's surrounding context. There is no requirement on the number of tunable tracking parameters as well as their mutual dependence in the learning process. However, there is a need of training data which should be representative enough to make this algorithm generic. The third approach takes full advantage of features (hand-crafted and learned features) and tracklet affinity measurements proposed for the Re-id task and adapting them to MOT. Framework can work with or without training step depending on the tracklet affinity measurement.

The experiments over three datasets, MOT2015, MOT2017 and ParkingLot show that the third approach is the most effective. The first and the third (without training) approaches are the most generic while the third approach (with training) necessitates the most supervision. Therefore, depending on the application as well as the availability of a training dataset, the most appropriate MOT algorithm could be selected.

---

**Keywords :** MOT, people tracking

# ACKNOWLEDGMENTS

---

---

I would like to thank Dr. Jean-Marc ODOBEZ, from IDIAP Research Institute, Switzerland, Prof. Jordi GONZALEZ from ISELab of Barcelona University and Prof. Serge MIGUET from ICOM, Universite Lumiere Lyon 2, France , for accepting to review my PhD manuscript and for their pertinent feedbacks. I also would like to give my thanks to Prof. Precioso FREDERIC - I3S - Nice University, France for accepting to be the president of the committee.

I sincerely thank my thesis supervisors Francois BREMOND for what they have done for me. It is my great chance to work with them. Thanks for teaching me how to communicate with the scientific community, for being very patient to repeat the scientific explanations several times due to my limitations on knowledge and foreign language. His high requirements have helped me to obtain significant progress in my research capacity. He guided me the necessary skills to express and formalize the scientific ideas. Thanks for giving me a lot of new ideas to improve my thesis. I am sorry not to be a good enough student to understand quickly and explore all these ideas in this manuscript. With his availability and kindness, he has taught me the necessary scientific and technical knowledge as well as redaction aspects for my PhD study. He also gave me all necessary supports so that I could complete this thesis. I have also learned from him how to face to the difficult situations and how important the human relationship is. I really appreciate him.

I then would like to acknowledge Jane for helping me to solve a lot of complex administrative and official problems that I never imagine.

Many special thanks are also to all of my colleagues in the STARS team for their kindness as well as their scientific and technical supports during my thesis period, especially Duc-Phu, Etienne, Julien, Farhood, Furqan, Javier, Hung, Carlos, Annie. All of them have given me a very warm and friendly working environment.

Big thanks are to my Vietnamese friends for helping me to overcome my homesickness. I will always keep in mind all good moments we have spent together.

I also appreciate my colleagues from the faculty of Information Technology of ThaiNguyen University of Information and Communication Technology ( ThaiNguyen city, Vietnam) who have given me the best conditions so that I could completely focus on my study in France. I sincerely thank Dr. Viet-Binh PHAM, director of the University, for his kindness and supports to my study plan. Thank researchers (Dr Thi-Lan LE, Dr Thi-Thanh-Hai NGUYEN, Dr Hai TRAN) at MICA institute (Hanoi, Vietnam) for instructing me the fundamental knowledge of Computer Vision which support me a lot to start my PhD study.

A big thank to my all family members, especially my mother, Thi-Thuyet HOANG, for their

full encouragements and perfect supports during my studies. It has been more than three years since I lived far from family. It does not count short or quick but still long enough for helping me to recognize how important my family is in my life.

The most special and greatest thanks are for my boyfriend, Ngoc-Huy VU. Thanks for supporting me entirely and perfectly all along my PhD study. Thanks for being always beside me and sharing with me all happy as well as hard moments. This thesis is thanks to him and is for him.

Finally, I would like to thank and to present my excuses to all the persons I have forgotten to mention in this section.

Thi-Lan-Anh NGUYEN  
thi-lan-anh.nguyen@sophia.inria.fr  
Sophia Antipolis, France

# CONTENTS

<b>Acknowledgements</b>	i
<b>Figures</b>	x
<b>Tables</b>	xii
<b>1 Introduction</b>	1
1.1 Multi-object tracking (MOT)	2
1.2 Motivations	3
1.3 Contributions	4
1.4 Thesis structure	6
<b>2 Multi-Object Tracking, A Literature Overview</b>	9
2.1 MOT categorization	10
2.1.1 Online tracking	10
2.1.2 Offline tracking	10
2.2 MOT models	11
2.2.1 Observation model	12
2.2.1.1 Appearance model	12
2.2.1.1.1 Features	12
2.2.1.1.2 Appearance model categories	14
2.2.1.2 Motion model	17
2.2.1.3 Exclusion model	19
2.2.1.4 Occlusion handling model	21
2.2.2 Association model	23
2.2.2.1 Probabilistic inference	23
2.2.2.2 Deterministic optimization	23
2.2.2.2.1 Local data association	24
2.2.2.2.2 Global data association	24
2.3 Trends in MOT	25

2.3.1	Data association	26
2.3.2	Affinity and appearance	26
2.3.3	Deep learning	26
2.4	Proposals	27
<b>3</b>	<b>General Definitions, Functions and MOT Evaluation</b>	<b>29</b>
3.1	Definitions	29
3.1.1	Tracklet	29
3.1.2	Candidates and Neighbours	30
3.2	Features	30
3.2.1	Node features	31
3.2.1.1	Individual features	32
3.2.1.2	Surrounding features	35
3.2.2	Tracklet features	37
3.3	Tracklet functions	37
3.3.1	Tracklet filtering	37
3.3.2	Interpolation	38
3.4	MOT Evaluation	38
3.4.1	Metrics	38
3.4.2	Datasets	39
3.4.3	Some evaluation issues	41
<b>4</b>	<b>Multi-Person Tracking based on an Online Estimation of Tracklet Feature Reliability</b>	<b>47</b>
<b>[80]</b>		<b>47</b>
4.1	Introduction	47
4.2	Related work	48
4.3	The proposed approach	49
4.3.1	The framework	50
4.3.2	Tracklet representation	51
4.3.3	Tracklet feature similarities	51
4.3.4	Feature weight computation	56
4.3.5	Tracklet linking	57
4.4	Evaluation	58
4.4.1	Performance evaluation	58
4.4.2	Tracking performance comparison	60
4.5	Conclusions	61



<b>5 Multi-Person Tracking Driven by Tracklet Surrounding Context [79]</b>	<b>65</b>
5.1 Introduction	65
5.2 Related work	66
5.3 The proposed framework	67
5.3.1 Video context	68
5.3.1.1 Codebook modeling of a video context	71
5.3.1.2 Context Distance	72
5.3.2 Tracklet features	73
5.3.3 Tracklet representation	74
5.3.4 Tracking parameter tuning	74
5.3.4.1 Hypothesis	74
5.3.4.2 Offline Tracking Parameter learning	75
5.3.4.3 Online Tracking Parameter tuning	76
5.3.4.4 Tracklet linking	77
5.4 Evaluation	77
5.4.1 Datasets	77
5.4.2 System parameters	78
5.4.3 Performance evaluation	78
5.4.3.1 PETs 2009 dataset	78
5.4.3.2 TUD dataset	79
5.4.3.3 Tracking performance comparison	80
5.5 Conclusions and future work	82
<b>6 Re-id based Multi-Person Tracking [81]</b>	<b>83</b>
6.1 Introduction	83
6.2 Related work	84
6.3 Hand-crafted feature based MOT framework	86
6.3.1 Tracklet representation	87
6.3.2 Learning mixture parameters	88
6.3.3 Similarity metric for tracklet representations	88
6.3.3.1 Metric learning	88
6.3.3.2 Tracklet representation similarity	91
6.4 Learned feature based framework	92
6.4.1 Modified-VGG16 based feature extractor	93
6.4.2 Tracklet representation	93
6.5 Data association	94
6.6 Experiments	94

6.6.1 Tracking feature comparison	94
6.6.2 Tracking performance comparison	96
6.7 Conclusions	97
<b>7 Experiment and Comparison</b>	<b>99</b>
7.1 Introduction	99
7.2 The best tracker selection	100
7.2.1 Comparison	100
7.3 The state-of-the-art tracker comparison	102
7.3.1 MOT15 dataset	102
7.3.1.1 System parameter setting	102
7.3.1.2 The proposed tracking performance	102
7.3.1.3 The state-of-the-art comparison	102
7.3.2 MOT17 dataset	106
7.3.2.1 System parameter setting	106
7.3.2.2 The proposed tracking performance	106
7.3.2.3 The state-of-the-art comparison	108
7.4 Conclusions	109
<b>8 Conclusions</b>	<b>119</b>
8.1 Conclusion	119
8.1.1 Contributions	121
8.1.2 Limitations	121
8.1.2.1 Theoretical limitations	121
8.1.2.2 Experimental limitations	122
8.2 Proposed tracker comparison	122
8.3 Future work	123
<b>9 Publications</b>	<b>125</b>