Candidatura FCT Bolsa de Doutoramento

Área Científica Principal: Ciências da Engenharia e Tecnologias

Área Científica Secundária: Engenharia Eletrotécnica, Eletrónica e Informática

Subárea: Informática

Painel de avaliação: Ciências da Computação e Informática / Computer Sciences and Informatics

**Designação do Programa de Doutoramento a frequentar**

ProDEI is the Doctoral Programme in Informatics Engineering of FEUP (Faculty of Engineering, University of Porto). The main goal of ProDEI is to promote excellence through research and development, for students in Informatics Engineering topics.

All the ACM-IEEE Computing Curricula main areas (Computer Science, Computing Engineering, Information Systems, Software Engineering, Information Technology) are accommodate in this Doctoral Program.

Emphasis is given to theoretical fundamentals but also including the specification, project, modeling, representation, distribution and exploration of computing systems, as well as information acquisition, representation and processing, with special attention given to the integration of those technologies in organizational environments.

**Título do Programa de Trabalhos:**

Deep Learning Methods for Object Re-Identification and Image Comprehension on Urban Scenarios

Keywords (4):

* Computer Vision
* Pattern Recognition
* Machine Learning ´
* Image Captioning

Data de início do programa de trabalhos: 03-03-2020

Work programme starting date Duração (meses): 36

Data de início pretendida para a bolsa: 01-09-2020

Fellowship starting date Duração (meses): 36

**Carta Motivação:**

I am writing to express my interest in pursuing a Doctoral degree in the field of Computer Science and Informatics Engineering, as it has always been my long-age ambition to become a successful researcher in the field of Computer Science.

Thanks to the developed work in my two previous Master's degrees and research activities, it is glaring that studying and doing research are endeavors which I would like to engage in even more. While studying for my Master's degrees in Electrical and Computer Engineering and Computer Science, both at the University of Porto, and performing research work, I developed a strong interest in the fields of Computer Vision and Machine Learning, and I always deepening my knowledge on new developments on this field.

I believe there is no better way to continue my carrier in the research area and in the fields of Computer Vision and Machine Learning than the pursuit of a Doctoral degree. Considering my previous background as well the content attained during my both Master degrees at University of Porto, combined with the knowledge I have gathered from my previous studies and research work, I am confident that the pursuit of a Doctoral degree brings me a step closer to my goal of becoming a successful researcher in the novel fields of Computer Vision and Machine Learning.

I believe that I am a very diligent and highly motivated student and researcher: while studying for my Master degrees, I achieved relevant results in complementary studies, mainly in research works developed simultaneously. I have always worked to accomplish my goals and gain more knowledge in the Computer Science field. I have a strong background on the fields of Computer Vision and Machine Learning, and more recently on Deep Learning methods, namely by producing new solutions for my master thesis and research work. My developed work has culminated in the publication of articles on peer-review computer vision and deep learning related conferences, particularly related to Robust Detection and Tracking of Ground Vehicles using UAVs, and Lightweight Deep Learning Pipeline for Detection, Segmentation and Classification of Breast Cancer Anomalies.

Pursuing a Doctoral degree is an opportunity to which I would like to dedicate myself in the following years, and from where I will be able to contribute to the scientific community in the best way, by publishing relevant scientific contributions in peer-reviews conferences and journals of the field. Considering my academic background and working experience so far, and my desire to enrich mine and others' knowledge's in Computer Vision and Machine Learning, I am convinced that I will be a valuable addition to the scientific community since I am confident that I am capable of meeting the expectations.

**Sumário (max 150 palavras):**

Object Re-Identification (ReID) and Image Captioning aims to match objects, such as cars or persons, on an image scene acquired at different times or/and by non-overlapping cameras, and describe the objects, and related actions and interactions. In recent years this requirements have become increasingly popular in a wide range of applications, such as in public surveillance and city patterns learning and recognition. However, Re-ID and Image Captioning still poses challenging problems due to complex object’s poses, poor image illumination conditions, complex backgrounds, occlusion problems and inadequate camera viewpoints. These problematics impose serious limitations in the use of Re-ID and Image Caption algorithms, particularly, in public surveillance applications, requiring constant human supervision to extract useful information in order to make decisions in useful time.

This project aims to develop advanced machine learning based algorithms to automatically ReID objects in different urban scenes and reasoning, i.e. Image Caption, about the actions and interactions occurring on it. City services can be greatly improved if those common patterns and actions are learnt and then automatically identified, allowing the prompt mitigation to occurring threatening events in realtime, improving the city safety and quality.

**Estado da Arte (max 500 palavras):**

Object Re-identification (ReID) and Image Caption, or Scene Understanding, are fundamental requirements in modern urban surveillance systems, which still present several complex demands [22]. It is estimated that the major cities must handle thousands of events occurring simultaneously, with many of them occurring very fast and that requiring immediate actions from the authorities but are unnoticed by them [25].

Object ReID [21] is a problem that has been studied for several years by the computer vision community. It mainly consists of ReID of the different classes of objects that can appear at different instants or/and acquired by different cameras. Accurate object ReID and scene understanding are vital assets to automatically identify potential threats or unusual behaviors that are of interest for the authorities towards safer and better cities.

REDO:

However, in nowadays surveillance systems, the ReID of objects and scene undestanding are mostly performed by human operators, requiring constant focus on video feeds, in the search of suspicious behaviors and patterns. The operators are faced with the constant challenge of determining if a particular object reappears in another image in a universe of thousands of similar objects while simultaneously try to understand its behaviour. Any erroneous interpretation during critical events, such as robbery or car pursuits, may mislead authorities with wrong information, resulting in incorrect actions to be taken. Human object ReID and Scene Understanding requires vast amount of human concentration and are prone to error.

====

Efforts have been made to develop computational algorithms to automatically ReID objects acquired at different instants or/and by different cameras in the same scene or in a different scene, and extract image understatement using image captioning. However limitations are still to address regarding the combination of multiple objects recognition, ReID and Image Captioning into a generalized framework.

Earlier ReID algorithms are traditionally based on handcrafted features [23] or on Hybrid solutions combining feature fusion [24]. Recently, the proposed algorithms are based on Deep Learning (DL) approaches, with the most successful DL architecture being the Convolutional Neural Network (CNN) [14]. In the CNN, the data is propagated through layers via convolutions and other operations, e.g. pooling, flattening and dropout, giving to the network the ability to learn both local and high-level image features [9, 11]. Object ReID based on CNNs has already been proposed, with most of the current approaches exploring the use of pre-trained CNNs, by redefining the last layers to address the ReID problem [21], by integrating Siamese networks [4, 5, 11] and modified triplet loss [6, 7], and using spatial attention models [10, 12], novel augmentation techniques [13] and, lately, 3D-CNNs initially trained for action recognition [8], and Generative Adversarial Networks [15].

As to image captioning, which has as the main objective the extraction of textual description from images, most of the proposed solutions are based on CNNs for feature extraction [13, 20], combined with Recurrent Neural Networks (RNN) [16, 17] for natural language processing [17], by exploring LSTM models [19], Dependency Tree Relation (DTR) [17], and attention mechanisms [18]. However, the optimal architecture for each application has not yet been established and many opportunities are still open; in particular, image understatement has not been addressed properly to operate in multi-camera distributed environments. There is a bright future for DL applied to urban surveillance systems, with its huge potential to complement object ReID and Image Captioning into a single solution still unexplored.

**Objectivos (max 300 palavras):**

The principal objective of this PhD project is to develop and apply machine learning algorithms to automatically Re-Identify (ReID) objects in different scenes and obtain the reasoning about their behavior, which are vital tasks in modern urban surveillance systems. Specifically, the idea consists of working with multiple live video streams from urban scenarios, where the diversity of objects to be automatically identified and re-identified ranges from automobiles, trucks, persons, lost children's and elders, while simultaneously identify abnormal behaviors such as robberies or unusual crows concentrations. The input data will be color images from video surveillance feeds properly anonymized, allowing the extraction of useful information from the objects in the image scene and automatically identify potential threats.

The underlying purpose is to decrease the response time of the authorities, by providing relevant alerts to occurrences that require their prompt response while, at the same time, freeing the surveillance professionals and city authorities to mitigate other issues that concern the urban scenario operation.

The object ReID and scene understanding underlies multiple variables, that are hard to identify in useful time and which require the constant attention from a vast group of personal. The solutions to be developed are going to be useful to reduce the personal’s workload while providing information about city patterns that can be explored to improve the quality and safety of the citizens' life.

**Descrição Detalhada (max 1000 palavras):**

This project involves a set of tasks organized in distinct phases. The expected phases are consecutive, with some overlap between them. If needed, some of the phases may be revisited later to be updated, for example, due to the continuous development of the state of the art.

Phase I: review of the current literature - this has two fundamental objectives. First, to study the basic concepts, techniques and methods that have been developed and are currently in use, identifying their strengths and weaknesses as well as beginning to understand how each can be applied to this project. Two objectives will be addressed: i) to know how object Re-Identification (ReID) and Image Captioning are performed nowadays, and how computer vision and machine learning algorithms can increase the performance of urban surveillance systems, while reducing the number of false positives; and ii) to identify approaches based on machine learning and deep learning techniques that have been successfully used in object ReID and image captioning.

Importantly, a collection of bibliographic references will be gathered to serve as the basis for the development of the following phase. This will be continuously updated during the project, considering the fast developments occurring in this area.

Phase II: selecting and gathering data - one of the most important aspects of computer vision and machine learning areas is the quality and variability of the data used to train and validate the developed algorithms. Thus, this phase of this project will be dedicated to defining the details of the required data. First, a decision will be made about which databases to use. Different public databases are available, ranging from those containing pairs of image objects gathered under different light conditions and poses that are essential for object ReID, and Image Captioning databases containing textual descriptions about the objects and actions occurring on the images, which provides invaluable information for, for example, human action recognition and, in particular, for image based scene understanding. Both of kinds of databases will be explored, giving special attention to operate on data that was previously anonymized, since models should not rely on, for example, in biometric features to comply with Person Privacy regulation in practice.

Phase III: development and optimization of algorithms for ReID and Image Captioning - this will be the most complex and laborious phase, as several techniques and approaches will be researched and attempted, to develop and implement techniques to increase the accuracy in object ReID and Image Captioning. This involves a continuous back and forth process of fine-tuning specific aspects of the proposed ReID and Image Captioning algorithms and combine different tools to achieve the desired accuracy in different scenarios of operation. Whatever the implemented algorithms, they will be trained on part of the data previously collected through supervised learning. The validation dataset will then be used to perform a preliminary evaluation of the models under development, before tuning their hyperparameters and going back to the training set. In this phase, techniques of transfer learning will mostly be used, mainly to facilitate the use and improve the generalization of the developed deep learning models. On the other hand, to improve the robustness of the ReID and Image Captioning methods, high and low level features extracted from the imaged objects and new fusion schemes such as multi-modal, attention, novel model losses, and other variations of network architectures, will be employed. This information will be gathered and merged into a single framework to be deployed in a test scenario to be continuously updated for objects ReID and Image Captioning during the project.

Phase IV: testing the proposed ReID and Image Captioning methods and analysis of their results - At this point, the test dataset will be input into the models and the output compared with the state of the art benchmarks. To assess the performance, mainly the accuracy, of the proposed algorithms for object ReID and Image Captioning, statistical methods will be used: the Rank-1 Cumulative Matching Characteristic curve (CMC), mean Average Precision (mAP) and Bilingual Evaluation Understudy (BLEU). Other evaluation approaches that might prove useful may be considered later. To allow for a direct comparison of different algorithms for object ReID and Image Captioning, several challenges are organized every year using the same datasets for all algorithms under comparison. We plan to submit the developed algorithm and models to some of these challenges as, for example, the AI CITY CHALLENGE (www.aicitychallenge.org), which will enable to evaluate the object ReID algorithms that automatically identify unequivocally the same objects in images acquired by different cameras in the same scene. As to the automatic image captioning, we plan to evaluate our solutions on challenges that are occurring on the major computer vision and machine learning conferences such as Conceptual Captions organized by CVPR.

Phase V: writing of articles and of the thesis - an important phase of this project consists in writing texts which will constitute the written outcomes of this PhD. Particularly, the PhD thesis will be composed of several chapters populated in part by documents produced during the aforementioned phases of this project, manly scientific articles to be submitted and published in conference proceedings and peer reviewed scientific journals.

**Adequação da Equipa de Orientação e da Instituição de Acolhimento (max 300 palavras):**

This doctoral thesis project will be developed at INEGI - Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, which is part of LAETA - Laboratório Associado de Energia, Transportes e Aeronáutica, under the umbrella of FEUP - Faculdade de Engenharia da Universidade do Porto, which will ultimately award the PhD degree under the scope of the Doctoral Programme in Informatics Engineering.

INEGI’s research activity is integrated into the Portuguese network of research funded by Portuguese Foundation for Science and Technology (FCT) and has consistently been evaluated as excellent. In the last years, INEGI (www.inegi.up.pt) has given special attention to areas dedicated to Computer Vision and Machine learning. Professor João Tavares, the supervisor of this PhD project, has been deeply involved in these areas with several MSc, PhD and Post-Doc students that have been working on image processing and analysis, machine learning and biomechanics.

In addition, Professor João Tavares has organised several international scientific events related to those topics and produced many scientific publications (650 articles, 180 in ISI journals with 60 in ISI Q1 journals, 55 books, 50 book chapters, and 3 international patents). He is also founder and editor-in-chief of the journal “Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization” published by Taylor & Francis, editor-in-chief of the journal “Computer Methods in Biomechanics and Biomedical Engineering” published by Taylor & Francis, and co-founder and co-editor of the book series “Lecture Notes in Computational Vision and Biomechanics” published by Springer.

In the last years, he has established good working connections with medical and computer vision institutions such as Lenitudes Medical Center and Research as well as Bosch-Group Bosch Center for Artificial Intelligence, S.A, from which images and new problems will be gathered and addressed in the context of this project, while making use of an NVIDIA DGX GPU server available in the research group for training and validating the proposed algorithms.

**Cronograma**

Ficheiro excel em anexo.

**Referências Bibliográficas (max 30)**

[1] João Paulo Papa, Willian Paraguassu Amorim, Alexandre Xavier Falcão and João Manuel R. S. Tavares. Recent Advances on Optimum-Path Forest for Data Classification: Supervised, Semi-Supervised and Unsupervised Learning, Handbook of Pattern Recognition and Computer Vision, 5th edition, Chapter 1.6, DOI: 10.1142/9789814656535\_0006, ISBN: 978-981-4656-52-8, World Scientific, pp. 109-123, 2016.

[2] Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, Joao Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, Anant Madabhushi (editors). Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. ISSN: 978-3-030-00889-5, DOI: 10.1007/978-3-030-00889-5, Springer, 2018.

[3] João F. Nunes, Pedro M. Moreira and João Manuel R. S. Tavares. Human Motion Analysis and Simulation Tools: A Survey, Handbook of Research on Computational Simulation and Modeling in Engineering, Chapter 12, ISBN: 1466688238, DOI: 10.4018/978-1-4666-8823-0.ch012, IGI Global, pp. 356-386, 2015.

[4] Chen Shen, Zhongming Jin, Yiru Zhao, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and XianSheng Hua. Deep siamese network with multi-level similarity perception for person reidentification. In Proceedings of the 25th ACM international conference on Multimedia, pages 1942–1950, 2017.

[5] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE conference on computer vision and

pattern recognition, pp. 3908–3916, 2015.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person reidentification by multichannel parts-based CNN with improved triplet loss function. In

Proceedings of the iEEE conference on computer vision and pattern recognition, pp

1335–1344, 2016.

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

[8] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-based person reidentification via 3d convolutional networks and non-local attention. In Asian Conference

on Computer Vision, pp. 620–634. Springer, 2018.

[9] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724, 2017.

[10] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person reidentification by transfer learning of spatial-temporal patterns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7948–7956, 2018.

[11] VV Molchanov, BV Vishnyakov, YV Vizilter, OV Vishnyakova, and VA Knyaz. Pedestrian detection in video surveillance using fully convolutional yolo neural network. In Automated Visual Inspection and Machine Vision II, volume 10334, pp. 103340Q. International Society for Optics and Photonics, 2017.

[12] [Ehsan Yaghoubi, Diana Borza, Pendar Alirezazadeh, Aruna Kumar, and Hugo Proença. An implicit attention mechanism for deep learning pedestrian re-identification frameworks. arXiv preprint arXiv:2001.11267, 2020.

[13] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for

image recognition. In Proceedings of the IEEE conference on computer vision and pattern

recognition, pp. 770–778, 2016.

[15] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In Advances in neural information processing systems, pp 1222–1233, 2018.

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic

embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539,

2014b.

[17] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for

bidirectional image sentence mapping. In Advances in neural information processing

systems, pp. 1889–1897, 2014.

[18] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1242–1250, 2017.

[19] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651–4659, 2016.

[20] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In International conference on machine learning, pp. 595–603, 2014a.

[21] Chen, Weihua, et al. "Beyond triplet loss: a deep quadruplet network for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[22] Samatas, Minas. "Studying surveillance in Greece: Methodological and other problems related to an authoritarian surveillance culture." Surveillance & Society 3.2/3 (2005).

[23] Gao, Mu, Haizhou Ai, and Bo Bai. "A feature fusion strategy for person re-identification." 2016 IEEE international conference on image processing (ICIP). IEEE, 2016.

[24] Eisenbach, M., Kolarow, A., Vorndran, A., Niebling, J., Gross, H.M.: Evaluation of multi feature fusion at score level for appearance-based person re-identification. In: 2015 International Joint Conference on Neural Networks

(IJCNN), pp. 1–8. IEEE (2015)