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) can be 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 paid 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'm writing to express my interest in pursuing a Doctoral degree in the field com 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 I would like to engage in even more. While studying for my Master's degrees in Electronic and Computer Engineering and Computer Science, both at and the University of Porto, and performing research work, I developed a strong interest in the fields of Computer Vision and Machine Learning, Data Science, 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 the pedigree as well the content 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 degree, I achieved relevant results complementing studies with research work simultaneously. I have always worked to accomplish my goals and gain more knowledge in the Computer Science field. I have a strong background the the fields of Computer Vision and Machine Learning and more recently in Deep Learning methods, namely by producing works for master thesis and research work, that culminated in the publication of articles on peer-review computer vision and deep learning related conferences, with some published works regarding the Robust Detection and Tracking of Ground Vehicles using UAVs, and Lightweight Deep Learning Pipeline for Detection, Segmentation and Classification of Breast Cancer Anomalies. As relevant document I put focus of the article Lightweight Deep Learning Pipeline for Detection, Segmentation and Classification of Breast Cancer Anomalies, since it was the summary of one of my Master thesis that explores deep learning methodologies that are the main core of my current Phd proposal.

Pursuing a Doctoral degree is an opportunity 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'm confident that I am capable of meeting the expectations.

**Sumário (max 150 palavras):**

Object Re-Identification (ReID) and Image Captioning aims to match objects (cars, persons) in different places (time) using non-overlapping cameras and describe objects actions or interactions on an image scene. In recent years this requirements have become increasingly popular in a wide range of applications, such as public security, a criminal investigation, surveillance and city pattern recognition. However, ReID and Image Captioning still poses challenging problems due to object pose, image illumination, complex background, occluded region and camera viewpoints. These problematics imposes serious limitations in the use of Re-ID and Image captioning algorithms in public surveillance systems, requiring constant human supervision to extract useful information to make decisions in useful time.

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

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

As a result of the growth of the urban population worldwide [1], cities are consolidating their position as one of the central structures in human organizations. This concentration of resources and services around cities offers new opportunities to be exploited. Smart cities [2,3] are emerging as a paradigm to take advantage of these opportunities to improve their citizens’ lives. Smart cities use the sensing architecture deployed in the city to provide new and disruptive city-wide services both to the citizens and the policy-makers. One example of this is the use of video surveillance systems to comprehend city patterns. One of the main requirements on video systems concerns Object Re-identification (ReID) and Scene Understanding problematics, that at present time still posed different problems[22]. It is estimated that the major cities must handle thousands of events occurring simultaneously, with rapid occurring events that require immediate action passing unnoticed by authorities [25].

Object ReID[21] and Scene Understanding are major problems that has been subject to studies from several years by the computer vision community. It mainly consists of ReID of the different classes of objects that can appear in different camera and reasoning about their behaviours and actions. Accurate object ReID and scene understanding are vital assets to automatically identify potential threats or unusual behaviors that are of interest for authorities.

However, in nowadays surveillance, object ReID and scene undestanding are mostly performed by a team of human operators, requiring constant focus on video feeds, in the search of suspicious behaviors and patterns. 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 understand their 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 in different cameras from same and different scene, and extract image understatement using image captioning. However limitations are still to address regarding the combination of multiple object recognition, ReID and Image Captioning into a generalized single framework.

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

For image captioning, which has as the main objective the extraction of textual description of an image, with most authors relying on CNN for feature extraction[13, 20] in combination with Recurrent Neural Networks (RNN)[16, 17] used on natural language processing[17], by exploring LSTM[19] models, 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 a multi-camera distributed environment. There is a bright future for DL applied to urban surveillance systems and the potential to complement object ReID and Image Captioning into a single solution is still unexplored.

**Objectivos (max 300 palavras):**

The principal objective of this PhD project is to develop and apply machine learning tools to automatically Re-Identify (ReID) objects in different scenes and reasoning about their behavior, which are vital tasks for modern urban surveillance systems. Specifically, the plan 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 normal and abnormal behaviors such as robberies or unusual crows concentrations as well as understanding scene type or location, object properties, and their interactions.

The input data will be color images from video surveillance feeds properly anonymized, allowing the extraction of useful information from objects 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 surveillance professionals and city authorities to mitigate other issues that concern the urban scenario operation.

The object ReID and scene understatement underlies multiple variables, that are hard to identify in useful time and require their constant attention by a vast group of personal. The tools to be developed are going to be usuful to reduce the workload while at the same time providing information about city patterns that can be explored to improve the quality of the of the citizens' life.

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

This project involves a set of tasks organized in distinct phases. Although these are consecutive, some overlap is to be expected. Some phases may be revisited later to be updated as needed given new unexpected findings.

Phase I: a review of the current literature - this has two fundamental objectives. First, to study the basic concepts, techniques, and methods that have been developed and that are currently in use, identifying their strengths and weaknesses as well as beginning to understand how each can be applied to this project. Two domains will be explored: i) how object Re-Identification (ReID) and Image Captioning is 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) machine learning and deep learning techniques approach 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 in continued update thought the project, considering the fast-paced development in this area.

Phase II: selecting and gathering data - one of the most important aspects of computer vision and machine learning is the quality and variability of the data to train and to validate the algorithms. Thus, this phase of this project will be dedicated to defining the details of the data required. First, the choice of which databases to use. Different public databases are available, ranging from those containing pairs of image objects gathered in different light condition and pose that is essential for object ReID, and Image Captioning databases containing text description of objects and actions occurring on those images, providing invaluable additional information for human action recognition, and in particular, image scene understanding. Both databases genres will be explored, taking special care to operate on data that has been 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 the algorithms for ReID and Image Captioning - this will arguably be the most complex and laborious phase, as several techniques and approaches will be researched and attempted, to find and develop techniques to increase object ReID and Image Captioning accuracy. This involves a continuous back and forth process of fine-tuning specific aspects of the proposed ReID and Image Captioning algorithms and combines different tools to achieve the desired accuracy in different scenarios of operation. Whatever the algorithms implemented, they will be trained on part of the data previously collected through supervised learning. The validation set will then be used to perform a preliminary evaluation of the models under development, before tuning the 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 architectures. On the other hand, to improve the robustness of the ReID and Image Captioning methods, high and low level extracted features from objects image 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 the object ReID and Image Captioning during the future.

Phase IV: testing the proposed ReID and Image Captioning, and analysis of results - At this point, the test set will be input into the models and the output compared with the state of the art benchmarks. To determine the accuracy of object ReID and Image Captioning and assess the performance of the proposed algorithms, statistical methods will be utilized: the Rank-1 Cumulative Matching Characteristic curve (CMC), mean Average Precision (mAP), and Bilingual Evaluation Understudy (BLEU). Other evaluation methods 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 which use the same dataset for all algorithms. We plan on submitting the developed algorithm and models to some challenges, as for example the AI CITY CHALLENGE (www.aicitychallenge.org), enabling us to evaluate object ReID algorithms that automatically identify unequivocally the same objects in images acquired by different cameras in the same scene. For the automatic image captioning, we plan to evaluate our work on challenges that are occurring on the major computer vision and machine learning conferences such as Conceptual Captions organized by CVPR. Although those challenges were held with participants during the 2018 and 2019 editions of the CVPR conference, the challenge is still ongoing online and future editions are still planned.

Phase V: writing articles and the thesis - an important phase of this project consists in writing the text which will constitute the written outcomes of this Ph.D. Particularly, the Ph.D. thesis will be composed of several chapters populated in part by documents produced during the previous phases of the 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 the 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 Ph.D. project, has been deeply involved in these areas with several MSc, Ph.D. and Post-Doc students that worked on image processing and analysis, machine learning, and biomechanics.

In addition, Professor João Tavares has organized several international scientific events related to these topics and produced many scientific publications (650 articles, 170 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] United Nations. World Urbanization Prospects: The 2014 Revision, Highlights; Department of Economic and Social Affairs, Population Division, United Nations: New York, NY, USA, 2014.

[2]. Caragliu, A.; Del Bo, C.; Nijkamp, P. Smart cities in Europe. J. Urban Technol. 2011, 18, 65–82.

[3]. Shapiro, J.M. Smart cities: Quality of life, productivity, and the growth effects of human capital.Rev. Econ. Stat. 2006, 88, 324–335. DOI: 10.4018/978-1-4666-8823-0.ch012, IGI Global, pp. 356-386, September 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 forperson re-identification. In Proceedings of the IEEE conference on computer vision and

pattern recognition, pages 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, pages

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, pages 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, pages 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, page 103340Q. International Society for Optics and Photonics, 2017

[12] [Ehsan Yaghoubi, Diana Borza, Pendar Alirezazadeh, Aruna Kumar, and Hugo Proen¸ca. 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, pages 0–0, 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, pages 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, pages 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, pages 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, pages 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, pages 4651–4659, 2016

[20] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In International conference on machine learning, pages 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 scorelevel for appearance-based person re-identification. In: 2015 International Joint Conference on Neural Networks

(IJCNN), pp. 1–8. IEEE (2015)