Text-based Person re-Identification for human tracking

Yuya Takagi

Master's thesis

Master's Programme in Imaging and Light in Extended Reality (IMLEX)

School of Computing

University of Eastern Finland

February 2024

UNIVERSITY OF EASTERN FINLAND
Faculty of Science, Forestry and Technology
Joensuu
School of Computing
Master's Programme in Imaging and Light in Extended Reality (IMLEX)

Takagi, Yuya: Text-based Person re-Identification for human tracking
Master's thesis, 15 p.
Supervisor: Jun Miura
February 2024

Abstract: Person re-identification is used to

Keywords: Text-based person Re-identification; Textt-based person retrieval

Contents

2 Introduction **2**

2.1 Motivation 2

2.2 Objectives 2

2.3 Stakeholders 2

2.4 Document structure 2

1. Introduction

Simultaneous Localization and Mapping (SLAM) using LiDAR technology stands as a cornerstone in autonomous navigation systems, enabling real-time mapping and localization essential for the robust and safe operation of various autonomous platforms. However, difficulties arise in mapping environments that contain specular or transparent surfaces as laser rays' reflections and transmissions on these areas lead to inaccuracies in the generated map, posing potential hazards during navigation. Recent research has introduced several novel techniques aimed at addressing this issue, but these approaches often exhibit constraints: some rely on incident angles closely aligned to normal; others are limited by specialized material handling, lacking adaptability across diverse surfaces; certain algorithms struggle with real-time processing, impeding their practical application.

1.1 Motivation

1.2 Objectives

1.3 Stakeholders

1.4 Document structure

2. Introduction

Simultaneous Localization and Mapping (SLAM) using LiDAR technology stands as a cornerstone in autonomous navigation systems, enabling real-time mapping and localization essential for the robust and safe operation of various autonomous platforms. However, difficulties arise in mapping environments that contain specular or transparent surfaces as laser rays' reflections and transmissions on these areas lead to inaccuracies in the generated map, posing potential hazards during navigation. Recent research has introduced several novel techniques aimed at addressing this issue, but these approaches often exhibit constraints: some rely on incident angles closely aligned to normal; others are limited by specialized material handling, lacking adaptability across diverse surfaces; certain algorithms struggle with real-time processing, impeding their practical application.

2.1 Motivation

2.2 Objectives

2.3 Stakeholders

2.4 Document structure

3. Literature Review

3.1 Research Methodology

WIP

In the methodology section, the we first delves into the existing literature, drawing from a paper accessible through the platform "Paper with Code." This platform typically provides research papers along with their associated code implementations. The chosen paper appears to be selected based on its prominence, likely measured by its reported accuracy or success in the field. Following the identification of the primary paper, the researcher conducts a thorough review of its content, focusing particularly on aspects related to methodology. This involves understanding the proposed techniques, algorithms, and approaches presented in the paper to achieve high accuracy in the context of text-based person searches. The aim is to comprehend the nuances of the existing methodology and identify the key factors contributing to its success. In addition to the primary paper, the researcher examines two other papers that exhibit a significant difference in accuracy. This comparative analysis is valuable for gaining insights into different approaches within the field. The choice of these additional papers may be strategic, aiming to capture diverse perspectives or methodologies, especially if there is a notable contrast in their reported accuracy metrics. The researcher likely scrutinizes the methodologies of these selected papers, comparing and contrasting them with the primary paper. This comparative analysis helps identify the strengths and weaknesses of different approaches, shedding light on potential areas of improvement or innovation for the current research. Overall, the methodology involves a comprehensive exploration of relevant literature, with a focus on the primary paper selected from "Paper with Code." The intent is to understand the methodologies employed in achieving high accuracies and to leverage insights from other papers with varying performance metrics.

However, if only paperwithcode is used, the information obtained is limited and biased. To eliminate this bias, we decided to use scopus to search a wider range of papers by keyword search.

Identification

The following research question was defined:

"Can lightweight models maintain detection accuracy in text-based person search?"

From this research question, four main keywords that sufficiently explain the topic were used: person retrieval and vision language pre-training. Furthermore, synonyms and related terms were associated to these keywords to form keyword groups as follows:

- person retrieval:
  - person;
  - person detection;
  - person search.
- vision language pre-training:
  - VLP;
  - text based;
  - text.

From the keywords, we had a keyword search on scopus from the search strings as follows:

- ("person retrieval" OR "person" OR "person detection" OR "person search") AND ("vision language pre-training" OR "VLP" OR "text based" OR "text").

The Scopus search yielded a total of 20170 documents. Within this result, we set the subject area to Computer Science, document type to article and conference paper, language to only english, and set the open access to all open access. With this filters, 862 articles were found.

Screening

Various factors were taken into account for the exclusion of documents:

1. problem and goal were too different (e.g., building new hardware, analysis of leaf reflectance);
2. not sufficiently related to this work (e.g., focused on hyperspectral);
3. duplicates that were not automatically detected and excluded.

3.2 Transformer

WIP

The Transformer model, introduced by Vaswani et al., 2023, is an encoder-decoder architecture featuring an attention mechanism. Initially devised to address the limitations of the Seq2Seq model, such as the challenges in grasping global nuances and the need for faster training, the Transformer model leverages attention mechanisms. These mechanisms, akin to the gating mechanism found in LSTM networks, prioritize important information while suppressing irrelevant details. By employing this approach, the model can effectively discern crucial information from the vast array of input data for processing.

Structure of the Transformer is shown figure3.1. The model is structured as encoder-decoder model, most competitive neural sequence transduction models are also encoder-decoder model, with encoder maps an input of symbol representation $X \ \ x_1, ..., x_n$ toa continuous representation $Z \ \ z_1, ..., z_n$. The decoder will take the given z to generate an output sequence of $y_1, ..., y_n$ of symbols one at a time.

3.3 Vision-language pre-training

WIP

Vision and Language(VL) is a major research area for the causality of Computer Vision and Natural Language Processing (NLP), which aims to effectively learn from multimodal data. Some of the great success of language model pre-training in NLP(RoBERTa(Liu

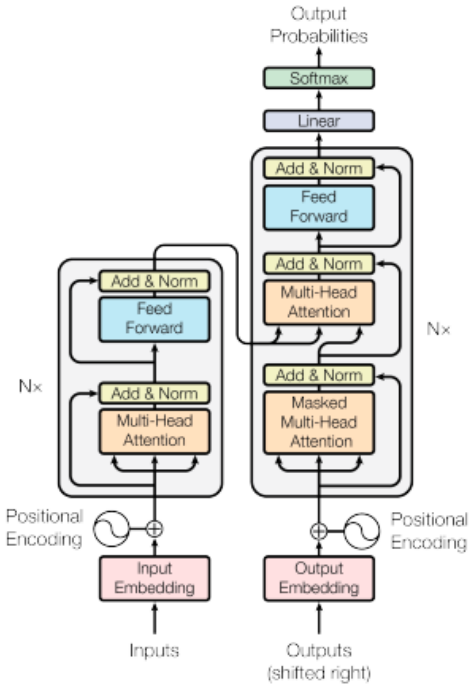

Figure 3.1: Structure of the Transformer

et al., 2020), BERT(Devlin et al., 2018), and GPT-3(Brown et al., 2020)) influenced the field of Vision-Language Pre-training (VLP) to grow attention from both fields.Towards this tasks, many studies have been proposed.

- VisualBERT A joint representation model for VL task which is designed to capture rich semantics in the image and associated text. The author integrated BERT for the NLP task with pretrained object proposal systems like Faster-RCNN. Figure3.2 shows how the model is trained.

- SimVLM Simple visual language model(SimVLM), introduced by Wang et al., 2022, is a minimalist pretraining framework for joint visual and textual representation. The motivation for this research is the limitations and complexity of existing vision-language pretraining models, which often require expensive annotations and multiple dataset-specific objectives. SimVLM aims to address these issues by utilizing large-scale weak supervision and training end-to-end with a single prefix language modeling objective.

- CLIP CLIP, introduced by Radford et al., 2021, is build on a vision transformer architecture, similar to BERT but instead of joint representation model, CLIP uses different encoder for both modal and from the extracted features, we calculate the

dot product. During the learning process, the diagonal component of this matrix are learned to have larger values than the others.

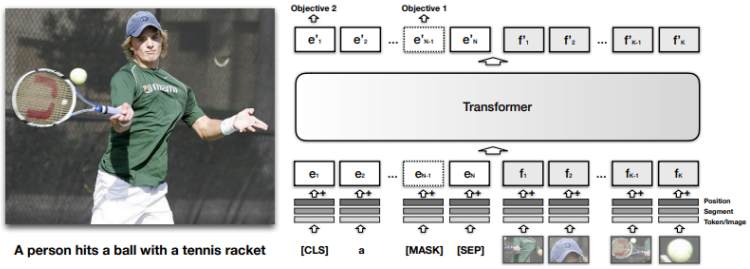

Figure 3.2: Representation of VisualBERT

3.4 Person Understanding Task

in this section we will talk about different method we can use for person understanding task.

3.4.1 Text-based re-identification

WIP

This section presents papers that study text-based person retrieval. A common issue addressed in each paper is the deficiency of the feature from text and image encoders. It has been confirmed that when the features of each modal are integrated, information is distorted or missing, which affects the accuracy of detection. Therefore, how to resolve this deficiency is key in this section.

- Relation and Sensitivity Aware representation learning
  This paper introduces a method called Relation and Sensitivity Aware representation learning (RaSa) that includes two novel tasks: Relation-Aware learning (RA) and Sensitivity-Aware learning (SA). It addresses the shortcomings of existing methods in text-based person search, where clustering representations of positive pairs without distinction leads to overfitting, particularly with weak positive pairs. Figure3.3 represents the overall structure of RaSa. RA mitigates overfitting by introducing a positive relation detection task to distinguish between strong and weak positive pairs. Additionally, the author emphasizes the common practice of learning invariant representation under data augmentation for robustness but goes

further by encouraging the representation to perceive sensitive transformations through SA, promoting enhanced robustness by detecting replaced words in textual descriptions.

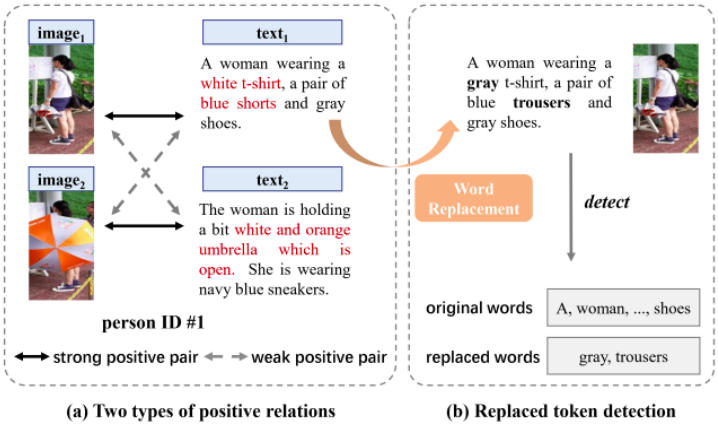

Figure 3.3: Structure of RaSa

- Implicit Relation Reasoning and Aligning

  The paper introduces a novel approach, called IRRA (Implicit Relation Reasoning and Aligning), for text-to-image person retrieval. This task involves identifying a person based on a given textual description. The main challenge is to establish an effective mapping between visual and textual modalities in a shared latent space. Unlike previous methods that use separately pre-trained unimodal models, IRRA addresses this challenge by introducing a cross-modal Implicit Relation Reasoning module. This module integrates visual cues into textual tokens through a masked language modeling paradigm, facilitating cross-modal interaction. To globally align visual and textual embeddings, the paper proposes Similarity Distribution Matching, which minimizes the KL divergence between image-text similarity distributions and normalized label matching distributions.

- Semantic-Aligned Feature Representation

  The paper(Li et al., 2021) focuses on text-based person search, aiming to retrieve images of a specific pedestrian based on a textual description. The primary challenge in this task is to bridge the inter-modality gap and align features across textual and visual modalities. The proposed solution is a semantic-aligned embedding method that automatically learns feature alignment between visual and textual representations. The method utilizes two Transformer-based backbones to encode robust feature representations for images and texts. Additionally, a semantic-aligned feature aggregation network is introduced, incorporating a

multi-head attention module constrained by a cross-modality part alignment loss and a diversity loss.

WIP

3.4.2 person attribute recognition

person attribute recognition has seeked to have supervision ... made a mask on the person to set the person ... tried to slice the image so that we can extract specific parts from the person ... tried to create a filter

- A Simple and Robust Correlation Filtering Suo et al., 2022
- Vision-Guided Semantic-Group Network He et al., 2023
- Pose-Guided Multi-Granularity Attention Network Jing et al., 2019

4. Design and implementation

4.1 Requirements specification

4.2 Software Architecture

4.3 Algorithm implementation

5. Evaluation

5.1 Test suite

5.2 Experiments

5.3 Results

6. Conclusions and outlook

Ignore this. It's for printing the glossary until I have other terms: mathematics

A. Prisma Automator

The “Prisma Automator” program was developed to automate the initial steps outlined in the PRISMA2020 Statement**prismastatement**. This process, typically done manually, involves formulating search strings, retrieving document metadata, and filtering results — tasks that become increasingly repetitive with more keyword combinations. The program aims to simplify user interaction by handling these steps, requiring only input of desired keywords and subsequent monitoring of the resulting document pool.

Comprising two classes, “Splitter” and “Collector”, Prisma Automator facilitates the generation of search strings (splits) and interacts with the Scopus API to retrieve, clean, and save results locally. Both classes offer streamlined functionality through the “split()” method in Splitter and the “run()” method in Collector, but users have the flexibility to employ other methods or customize functionality as needed.

Prisma Automator is an open-source project available at https://github.com/Fabulani/prisma-automator.

Glossary

mathematics Mathematics is what mathematicians do. 12

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

He, S., Luo, H., Jiang, W., Jiang, X., & Ding, H. (2023). Vgsg: Vision-guided semantic-group network for text-based person search.

Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., & Tan, T. (2019). Pose-guided multi-granularity attention network for text-based person search.

Li, S., Cao, M., & Zhang, M. (2021). Learning semantic-aligned feature representation for text-based person search.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2020). Ro{bert}a: A robustly optimized {bert} pretraining approach. https://openreview.net/forum?id=SyxS0T4tvS

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Suo, W., Sun, M., & Niu, e., Kai. (2022). A simple and robust correlation filtering method for text-based person search. *The European Conference on Computer Vision (ECCV)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2022). Simvlm: Simple visual language model pretraining with weak supervision.