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

January 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, 13 p.
Supervisor: Jun Miura
January 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:

"How can we discern and identify the details of a person 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

The Transformer was initially introduced for machine translation within the realm of natural language processing (NLP)Vaswani et al., 2023. Previous NLP models utilizes convolutional neural networks(CNN) or recurrent neural networks(RNN) for encoder and decoder. However, training this model is time consuming and required tremendous amount of labeled datasets. The transformer solves those problems by implementing an attention unit.

3.3 Vision-language pre-training

Vision language pre-training has drawn increasing attention after Transformer came up. Transformer is meant for NLP tasks, but the researcher found out that this model is capable of many tasks including information retrieval from the image, classification, etc. Since then, many follow-up works are proposed, such as BERT, CLIP, RoBERTa, GPT series. we know that bigger the transformer layer, the better accuracy. some methods like gpt4 has billions of parameters there are research that seeks into less parameters, enabling small model with efficient accuracies.

vision language like vibert albef CLIP

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

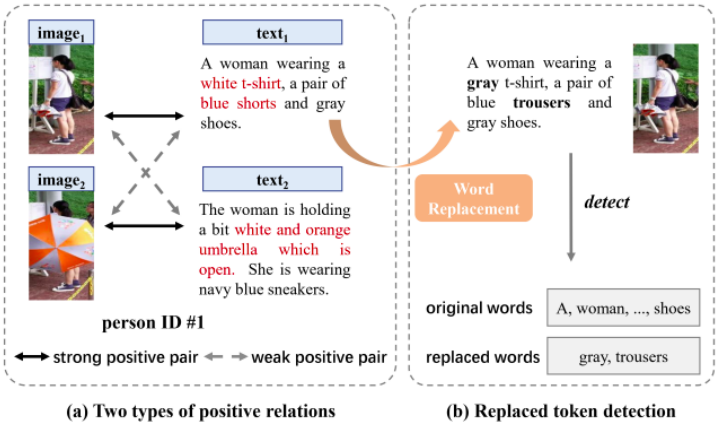
WIP

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.

RaSa: Relation and Sensitivity Aware e Representation Learning for Text-based Person Search

WIP

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. 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.



(a) Two types of positive relations (b) Replaced token detection

Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval

The paper introduces a novel approach, called IRRA (Implicit Relation Reasoning and Aligning), for text-to-image person retrieval. This task i'nvolves 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.

Learning Semantic-Aligned Feature Representation for Text-based Person Search

The paper 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.

3.4.2 Image-based re-identification

with the image based methods, we have the old fashion way of cnn

3.4.3 person attribute recognition

person attribute recognintion 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

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. 10

References

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