Video-Based Convolutional Attention for Person Re-Identification

Marco Zamprogno Marco Passon Niki Martinel Giuseppe Serra Giuseppe Lancioni Christian Micheloni Carlo Tasso Gian Luca Foresti

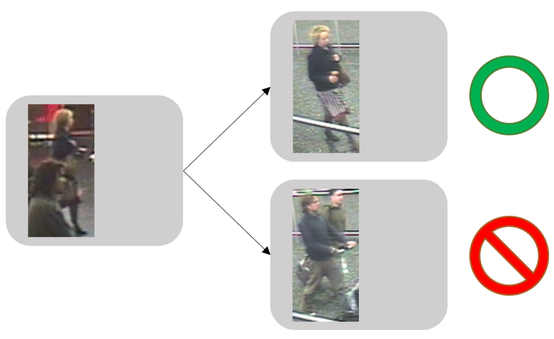
Università degli Studi di Udine, Udine (UD), Italy

Abstract

In this paper we consider the problem of video-based person re-identification, which is the task of associating videos of the same person captured by different and non-overlapping cameras. We introduce an attention mechanisms to capture the relevant information both at frame level (spatial information) and at video level (temporal information given by the importance of a specific frame within the sequence) at the same time, providing a very simple architecture while still achieving state-of-the-art performance.

Introduction

Re-Identification (ReID): given an image or video of a person taken from one camera, is the process of re-associating the person by analyzing images or videos taken from a different camera with non-overlapping field of view.



We propose an approach to the problem of video-based person re-identification that is characterized by two main aspects:

- a deep neural network architecture based on a Siamese framework [1] which evaluates the similarity of the query video to a candidate one;
- a novel spatio-temporal attention mechanism with the aim to select relevant information from different areas of the frames of the input video, and from their evolution over time.

The Proposed Approach

The proposed approach (see Fig. 1) is based on a Siamese network [1], which is composed by two identical networks:

- the first is fed with the query video;
- the second is fed instead with the candidate video to be compared.

Fig. 2 shows the proposed neural network. The main blocks are:

- the initial convolutional network;
- the Spatio-Temporal Attention Module;
- the final part which performs averaging and normalization.

The bottom part explodes the Attentive ConvLSTM and gives an idea of the multiple refining steps.

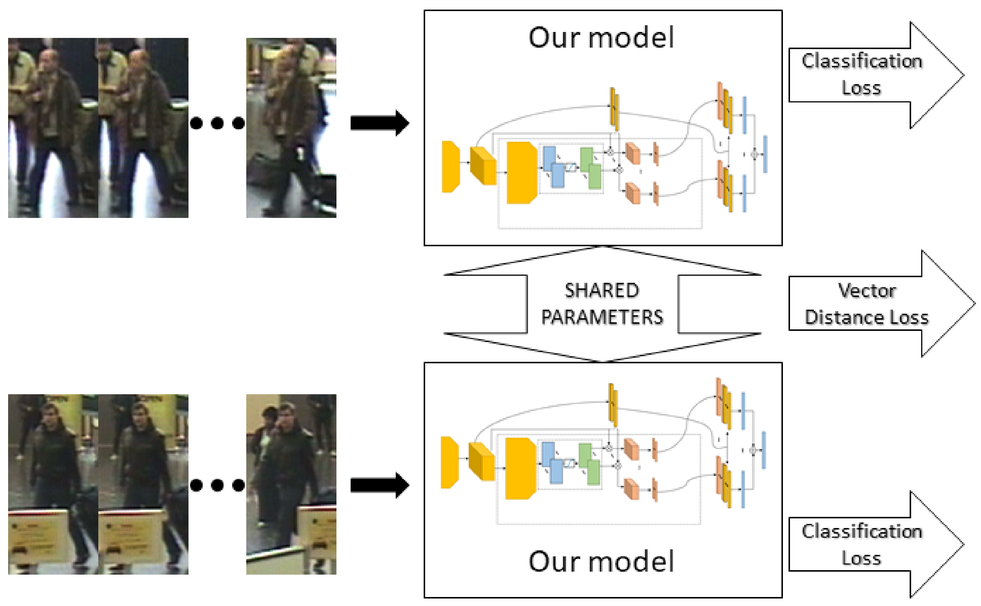


Figure 1: Siamese network scheme.

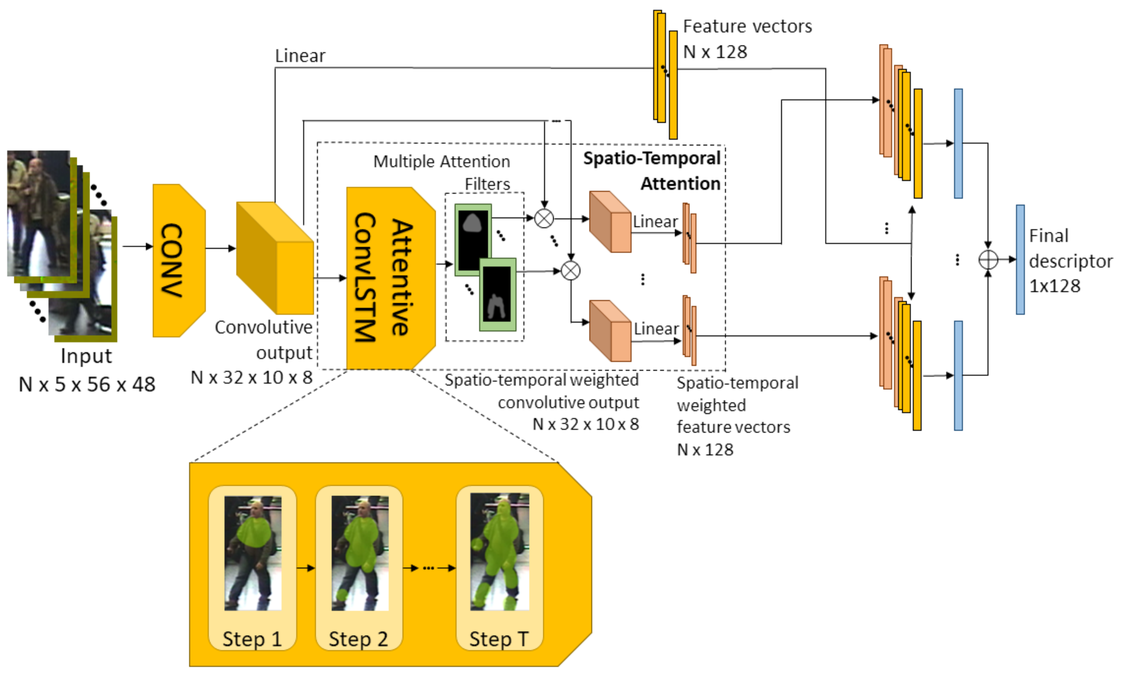


Figure 2: Our neural network model.

Dataset

To train and test our model we used the well-known challenging dataset iLIDS-VID, that contains many occlusions, severe illumination changes and background clutters. It consists of videos of 300 distinct people, and for each person there are two different video sequences, captured by two non-overlapping cameras.

Table 1: Comparison with state-of-the-art methods.

iLIDS-VID				
Methods	Rank-1	Rank-5	Rank-10	Rank-20
Ours	**63.3**	**87.4**	94	97.8
Xu	62	86	94	**98**
Zhang	60.2	85.1	-	94.2
McLaughlin	58	84	91	96
Zhengl	53	81.4	-	95.1
Yan	49.3	76.8	85.3	90.1

Results

- Despite beeing a simple architecture, our solution outperforms other methods proposed in the literature on 2 metrics out of 4 (Table 1).
- We also compared the results of our model when using different numbers of filters for the Spatio-Temporal Attention Module, with three providing the best performance (Table 2).

Table 2: Results obtained using an increasing number of filters.

Average results using different number of filters				
#filters	Rank-1	Rank-5	Rank-10	Rank-20
0	60.5	84.8	93	96.9
1	59.4	85.7	93.2	97.4
2	63	**87.7**	93.9	97.3
3	**63.3**	87.4	**94**	**97.8**
4	59.6	87.2	93.9	97.7

References

[1] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[2] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

Acknowledgements

This project was partially supported by the FVG P.O.R. FESR 2014-2020 fund, project "Design of a Digital Assistant based on machine learning and natural language, and by the "PREscriptive Situational awareness for cooperative autoorganizing aerial sensor NETworks" project CIG68827500FB.