Signal Processing 142 (2018) 104–113

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance

Wei Huang[a], Huijun Ding[b,*], Guang Chen[c]

[a] Department of Computer Science, School of Information Engineering, Nanchang University, Jiangxi, China
[b] Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Department of Biomedical Engineering, School of Medicine, Shenzhen University, Guangdong, China
[c] Xian Communications Institute, Shaanxi, China

ARTICLE INFO

Article history:
Received 22 February 2017
Revised 15 July 2017
Accepted 17 July 2017
Available online 18 July 2017

Keywords:
Deep metric learning
Deep residual networks
Localization
Video surveillance

ABSTRACT

Moving human localization is the first pre-requisite step of human activity analysis in video surveillance. Identifying human targets accurately and efficiently is always of high demands in computer vision studies. Also, learning is often indispensable in contemporary moving human localization, and unknown parameters of proposed methods need to be properly adjusted to guarantee the final localization performance. Such a task can be facilitated with the help of popular deep learning techniques, especially when enormous surveillance video clips become commonly seen nowadays. In this study, the metric learning problem in moving human localization is emphasized, and a new deep multi-channel residual networks-based metric learning method is introduced for the first time. Specifically, the deep metric learning problem in this new method is solved within a ranking procedure via both the conventional stochastic gradient descent algorithm and a more efficient proximal gradient descent algorithm. Comprehensive experiments are conducted and this new method is compared with several other popular deep learning-based approaches. Qualitative and quantitative analysis are conducted from the statistical perspective, to evaluate all localization outcomes obtained by all compared methods based on two specific measurements. The localization performance of this new method is suggested to be promising after the comprehensive analysis.

1. Introduction

Human activity analysis is an emerging and valuable topic in data mining and associated fields for the time being. When the topic is associated with different modalities of multi-media, human activity analysis is capable to bring about diverse tasks worthy of investigations. For example, the main purpose of human activity analysis in video surveillance is to discern an individual's behavior or a group's movements across consecutive frames of a surveillance video clip. When human activity analysis in social network is focused, the task then turns into the identification of specific ending users or the analysis of their connections among heterogeneous multi-media. When the recently popular smart-wearable devices are emphasized, human activity analysis is able to enclose a broad range of studies depending on specific characteristics and usages of particular devices. To sum up, it is necessary to clarify the role of human activity analysis in specific applications and propose effective methods to facilitate key problems in the human activity analysis.

In this study, human activity analysis in video surveillance is considered and its pre-requisite step of moving human localization across consecutive frames in a surveillance video clip is emphasized. The reason why moving human localization is essential for human activity analysis in video surveillance is easy to be perceived, as the human targets need to be localized first before actual analysis is conducted to analyze her/his behavior or activities. Generally speaking, moving human localization is a popular yet challenging task in pattern recognition and computer vision at the current stage, and there are many tough problems in moving human localization, including illumination changes, partial/full obstacles, rigid/non-rigid geometric appearance changes, etc. Although introducing a versatile method to tackle all challenging problems in surveillance video clips of large diversities still seems obscure at present, many researchers endeavor to introduce new models and algorithms to tackle specific problems. Early attempts to handle the moving human localization problem mainly rely on well-

* Corresponding author.
E-mail addresses: hjding@szu.edu.cn, huijun.d@gmail.com (H. Ding).

http://dx.doi.org/10.1016/j.sigpro.2017.07.015
0165-1684/

established pattern recognition models or machine learning techniques [1–5]. For instance, the popular boosting [1], support vector machine (SVM) [2], random forests [3] models are incorporated in bringing about various localization methods. Also, other sophisticated models, such as naive Bayes classifiers with fixed random basis [4], structured SVM [5], etc, are throughly investigated as well for the localization task. Some popular studies in recent years are introduced and discussed as follows. The conventional single-view learning strategy is analyzed in [6], and an alternative multi-view learning scheme based on intact space learning techniques is also introduced in this work. Encoded complementary information in multiple views is integrated to discover a latent intact representation of the utilized data, and the combination of multiple views of data is assumed to be capable to provide abundant information for a later learning. In [7], cognitive psychology principles are incorporated to design a flexible representation of moving objects whose shapes are likely to be changed, and a multi-store tracker consisting of long-term as well as short-term memories of appearances is introduced in this work. In [8], the multiple pedestrian tracking problem is emphasized, and a couple-states analysis including hidden states via a Markov chain transition process as well as latent states for semantic understandings is introduced in this work. It can be concluded from those aforementioned studies that, learning is indispensable in these up-to-date localization studies and it is often realized from the perspective of shallow learning. Also, after the localization task is well tackled, other high-level understandings, including the action recognition of each individual moving target or the activity interpretation of an ensemble of moving targets, can be fulfilled, thereafter [9,10].

Provided the fact that deep learning techniques receive much research attention beginning from the last few years, they have already been demonstrated to be more effective than many conventional shallow learning techniques within a broad range of applications, including audio recognition, digits recognition, image classification, etc [11]. It inspired us to solve the moving human localization problem in video surveillance from the perspective of state-of-the-art deep learning techniques. Therefore, in this study, the learning problem in moving human localization is tackled, and more specifically, the metric/metric-based similarity learning problem is emphasized. The reason why similarity learning is essential in moving human localization is easy to perceive. Generally speaking, the moving human target is normally considered as the foreground object, which needs to be discerned from the background before being localized in consecutive frames of a video clip. Therefore, a suitable similarity measure is necessary to group pixels belonging to this foreground object together while differentiate pixels between foreground and background, simultaneously. Moreover, determining such a suitable similarity needs to be driven by the video data via learnings. Since an ordinary metric-based similarity is often defined based on a typical metric, learning a metric-based similarity is equivalent to learning its corresponding metric. Thus, in order to avoid any ambiguity, the term metric learning will be utilized. In this study, a novel deep multi-channel residual networks-based metric learning method is introduced for the first time to realize moving human localization in video surveillance. The intuition of proposing such a new deep multi-channel residual network in metric learning is explained as follows. First, deep residual network is one of the latest deep learning models, which has demonstrated its outstanding performance of non-linear generalization in various studies in the last 2 years. Incorporating such a state-of-the-art deep learning method as the basic model is helpful for performance boosting of this study. Second, the raw video data in moving human localization can be extracted and represented in various forms (i.e., color space, texture space, optical flow, etc). A fusion of them in a single-channel deep residual network is clearly not proper and a multi-channel structure is necessary, which has not yet been proposed for metric/similarity learning till now. The organization of this paper is elaborated as follows. In Section 2, a comprehensive literature review of metric learning is provided. Metric learning from both the conventional shallow learning perspective and the currently popular deep learning point of view are discussed, with representative studies introduced. In Section 3, theoretical aspects of the new metric learning method are explained and two learning algorithms following a ranking model within this new method are derived. In Section 4, comprehensive experiments are conducted based on a large video database and the new method is compared with several other deep learning-based localization methods. Both qualitative and quantitative analysis are conducted based on all localization outcomes obtained by all methods, from the statistical point of view. In Section 5, the conclusion of this study is drawn and the potential future direction is suggested.

2. Related works

In psychology, there is an important theoretical assumption stating that, the similarity between two objects is inversely proportional towards the pairwise psychological distance between them. Hence, the more two objects are far away from each other, the less similar they are supposed to be. Therefore, the notion of similarity is closely related to the concept of distance. For metric, it is a special form of distance with particular characteristics. To be specific, a metric should be non-negative (i.e., the special form of distance between two objects should be equal to or more than zero). When the zero-value metric is obtained, it suggests that the two objects are overlapping or they are actually the same object. Moreover, the metric measured from object A to object B should be equivalent towards that from object B to object A (i.e., un-directional). Last but not least, the triangle inequality should be obeyed in metric as well. Based on such a special form of distance, when the similarity is proposed, it becomes the metric-based similarity.

Metric-based similarity is often of great importance in machine learning studies. For example, in the popular *k*-nearest-neighbor (KNN) classifier, the main idea is to classify an object into a group, which is the most common one among groups of its *k* nearest objects. Metric-based similarity in KNN is often used to indicate those nearest objects, and the performance of KNN is known to be highly influenced by such an incorporated metric-based similarity. In general image retrieval studies, it is also widely accepted that the metric-based similarity plays a crucial part since it helps to determine the degree of relevance of retrieved images to the query. In Section 2.1, the concept of metric and the notion of metric-based similarity is the focus. Characteristics of metrics are presented with several popular metrics introduced. In Section 2.2, popular metric / metric-based similarity learning methods proposed from both the conventional shallow learning perspective and the currently popular deep learning point of view are reviewed and discussed.

2.1. Metric and metric-based similarity

Metric is a special type of distance with several unique characteristics satisfied. Given three objects *X, Y*, and *Z; D* denotes the distance between any two pairwise objects. In order to make *D* a metric, four axioms must be satisfied simultaneously and they are presented as follows.

- Non-negativity: $D(X, Y) \geq 0$
- Identity of indiscernibles: $D(X, Y) = 0$ iff $X = Y$
- Symmetry: $D(X, Y) = D(Y, X)$
- Triangle inequality: $D(X, Y) \leq D(X, Z) + D(Y, Z)$

Violating one or more axioms can result in the named generalized metric, which in fact does not strictly follow all four requirements of becoming a metric, but can be considered as a spe-