Engineering Applications of Artificial Intelligence 65 (2017) 60–67

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Multi-shot human re-identification using a fast multi-scale video covariance descriptor

Bassem Hadjkacem [a,*], Walid Ayedi [a], Mohamed Abid [a], Hichem Snoussi [b]

[a] CES Research Laboratory, National Engineering School of Sfax, Sfax University, 3052 Sfax, Tunisia
[b] LM2S Research Laboratory, Charles Delaunay Institute (FRE CNRS 2848), University of Technology of Troyes, 10010, Troyes, France

ARTICLE INFO

Keywords:
Person re-identification
Multi-scale features
Video Tree Structure
Spatio-temporal covariance
Moving regions

ABSTRACT

Multi-shot person re-identification in non-overlapping camera networks has become an important research area. In order to tackle this problem, a robust and adaptive person modeling against occlusion and uncontrolled changes is required. In this paper, a new Multi-Scale Video Covariance (MS-VC) unsupervised approach was proposed to efficiently describe human in motion and requires no labeled training data. The MS-VC approach is based on the computing of the features extracted from a new structured representation called Video Tree Structure (VIDTREST) of any video sequence and can efficiently describe behavioral biometrics and appearance of each human by combining spatio-temporal information in a fixed-size vector. The VIDTREST model captures moving regions of interest. In addition, it decreases the color weight which can discard background noise and resolve clothing similarity cases in the appearance models and other changes. Furthermore, a fast algorithm was suggested to decompose each sequence under VIDTREST, extract its multi-scale features and compute its covariance matrices in one pass. The proposed method was evaluated with CAVIAR and PRID datasets. Our experimental results outperform the recognition rates of the existing unsupervised approaches in-the-state-of-the-art.

1. Introduction

Recently, perception in intelligent video surveillance systems has become more and more sophisticated to respond to a security need in large public or private spaces (Wang, 2013). Human re-identification (Re-ID) is an important application area for artificial intelligence (Bedagkar-Gala and Shah, 2014; Oliver et al., 2016; Chahlaa et al., 2017). Indeed, a person Re-ID system enables to track any person through different disjoint camera views (Black et al., 2004; Albiol et al., 2012; Perdomo et al., 2013). In non-overlapping camera network, a person's appearance usually varies across cameras due to occlusion, variation of illumination and pose etc. Therefore, a robust modeling is necessary to re-identify a given person. Appearance-based methods can be divided in two groups: mono-shot and multi-shot approaches. The first group uses one image per camera for each person obtained via tracking, without exploiting the temporal information. Most methods, such as that of Ayedi et al. (2012), Bingpeng et al. (2014), Wang et al. (2016), Xiaokai (2016) and Chen et al. (2016) combine spatial appearance features (e.g. intensity, color, gradient, etc.) to predict the identities of people and measure the similarity between signatures using a person's pair of images.

The second group, such as that of Gheissari et al. (2006) Bedagkar-Gala and Shah (2011) Bazzani et al. (2013) Bak and Bremond (2014) Li et al. (2015a, b) and Hadjkacem et al. (2016a, b) analyzes multiple images per camera for each person to build signatures for this person.

On-the-other-hand, the similarity of clothes colors and occlusion usually disturbs the Re-ID results using appearance-based method. Mono-shot appearance features are intrinsically limited due to these disturbances. Some works used the combination between the appearance features and the physical biometrics like the face (Selvam and Karruppiah, 2016) or behavioral biometrics like the gait (Zhao et al., 2007; Tafazzoli and Safabakhsh, 2010; Chattopadhyay et al., 2015; Xing et al., 2015).

Based partially on the works of Ayedi et al. (2012) and Hadjkacem et al. (2016b), this paper offers two main contributions. The first consists in introducing a Multi-Scale Video Covariance (MS-VC) unsupervised approach which describes the image sequences in multi-scale features and captures moving regions of interest to encode behavioral biometrics and appearance features using a new model called Video Tree Structure (VIDTREST). The second presents a novel fast algorithm to decompose

[*] Corresponding author.
E-mail addresses: bassem.hadjkacem.tn@ieee.org (B. Hadjkacem), ayediwalid@yahoo.fr (W. Ayedi), Mohamed.abid@rnu.enis.tn (M. Abid), hichem.snoussi@utt.fr (H. Snoussi).

http://dx.doi.org/10.1016/j.engappai.2017.07.010
Received 12 April 2017; Received in revised form 14 June 2017; Accepted 13 July 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

each image sequence under the form of a tree through the VIDTREST model and generate multi-scale features of MS-VC descriptor. This work is different to Ayedi et al. (2012) which treats only the images and cannot capture the moving regions in image sequences and the gait over time.

Moreover Hadjkacem et al. (2016b) introduce a video covariance descriptor which extracts the spatio-temporal features and the global correlation between video frames without capturing the moving regions of interest in temporal axis and modeling the video sequence as tree form.

The rest of this paper is organized as follows: Section 2, presents the related work on multi-shot human Re-ID. The MS-VC approach is presented in Section 3. Section 4 reports the experiment results on different datasets. Finally, Section 5 concludes this paper and gives some directions for previous work.

2. Related works

The multiple-shot approaches use many images of the same person in different poses to get a more informative signature. Several descriptors have been proposed to model objects in a way invariant to translation, scaling, rotation and illumination. The person re-identification process has been extensively studied with methods generally falling into two categories; supervised and unsupervised methods. First, supervised learning based methods such as the deep neuronal network that learn to map the raw features into a new space with greater discriminative power (McLaughlin et al., 2016). These methods require a large amount of labeled training image or video sequence pairs which severely limits their scalability and it is quite expensive to collect them. Secondly, the unsupervised methods employ invariant feature based methods that attempt to extract features that are both discriminative and invariant to environmental, lighting and viewpoint changes (Hadjkacem et al., 2016b). Moreover, the multiple-shot approaches can be classified into two methods. These approaches use groups of pictures of the same person in different poses to get a more informative signature. Several appearance descriptors such as local, global and by region have been proposed to model objects in a way invariant to translation, scaling, rotation and illumination.

In fact, Gheissari et al. (2006) proposed a spatio-temporal graph to reject contours that are considered unstable information over time. Then, a triangulated model person is used to manage the correspondence between the body parts. In Hamdoun et al., (2008), the Re-ID of individuals is performed using an optimized implementation of the SURF. Interest points are collected from movies. The interesting points are stored in a KD-tree to speed up the queries treatment. The allocation of models is carried out by a polling approach and a vote is added to each model containing the closest descriptor. The appearance-based method proposed by Bazzani et al. (2013) condenses a set of frames of any person into a Histogram Plus Epitome (HPE) descriptor. This descriptor embeds a global chromatic content via a histogram representation and local descriptions via the epitomic analysis. Wang et al. (2014) presented a model capable of simultaneously selecting and matching discriminative video fragments from unregulated pairs of image sequences of a person. This approach adopts the HOG spatio-temporal descriptor devised for action recognition. Thereafter, Hadjkacem et al. (2016b) integrated the temporal axis in the covariance descriptor proposed by Tuzel et al., (2006) to extract the spatio-temporal features.

On the other hand, some works used the combination between appearance descriptors and biometrics to improve the recognition rate. In fact, the model proposed by Bedagkar-Gala and Shah (2011) is a part-based spatio-temporal appearance model that combines facial features and color. Bingpeng et al. (2014) extracted biologically inspired features (BIF), computed the BIF similarity features taken at neighboring scales using a covariance descriptor and combined BIF and covariance method into a single vector. They also used a face verification step within the proposed descriptor. This proposal is not robust with the pose changes and low resolution in the general context of video surveillance applications. However, the approach in Kawai et al. (2012) merges the gait and the color features and requires an observation view regarding the gallery and the probe dataset to define the score-level fusion. It is an unrealistic method in video surveillance application. Some gait-based studies are based on examining the silhouette of a person over time. The primary problem with the silhouette based approaches is that they fundamentally entangle the body shape and the gait. Lombardi et al. (2013) introduced a gait representation that encodes the limb motions regardless of the body shape. Using the Kinect sensor, some works like (Andersson and Araujo, 2015) identify the people through the gait.

Based on the benefits of the covariance descriptor introduced by Tuzel et al. (2006) in person detection, our approach was adopted to combine the appearance features and behavioral biometrics. In fact, the covariance descriptor could be computed from any type of image and can also discard the noise effects (Bak and Bremond, 2014). It could encode information on the feature (color, texture...) variances inside the region, their correlation with one another as well as their spatial layout. Besides, it could produce a compact and a fixed representation by fusing different types of features. More recently, it was developed by Ayedi et al., (2012) and Bak and Bremond, (2014) for one-shot people Re-ID (Bingpeng et al., 2014) and by Bilinski and Bremond (2015) for the purpose of action recognition. Inspired by the-state-of-the-art, the multi-scale video covariance approach has been introduced to identify the people's appearance during a walking cycle, extract the correlation between video frames and encode the moving region of interest through the novel VIDTREST model.

3. Multi-scale video covariance descriptor

We presented our human Re-ID approach based on a new MS-VC descriptor. It is described in Fig. 1. The First step is to treat, resize and concatenate the person's images extracted from a video streaming by a tracking process carried prior to our work. A new video sequence was obtained as an input of Re-ID process. Then, the new VIDTREST model was applied to the obtained video sequence in order to capture the moving regions of interest (Section 3.1). During the decomposition process, the MS-VC features are extracted and the node's feature vectors and MS-VC matrices are calculated (Section 3.2). Thereafter, a fast algorithm is introduced to generate the multi-scale features of the VIDTREST and compute the covariance matrices (Section 3.3). Finally, comparisons between MS-VC matrices are achieved using the mathematical approach.

3.1. Video Tree Structure

The VIDTREST model is a new structured and flexible representation of video sequences to extract and save relevant multi-scale features. It can be used to handle positional information of video sequence features like the color and the correlation between the frames. A video sequence can be modeled via a VIDTREST by dividing it into T (number of frames) equal size quadrants, which would split recursively into four equal size quadrants, until a stopping condition is verified. The root node of the VIDTREST 'C' represents the whole video sequence where each frame quadrant is represented by a quadtree node 'c_t' storing any kind of information about the corresponding image quadrant. In this work, a covariance matrix is stored. If a video frame does not conform to the chosen split criteria, the root node will have four descendant nodes representing the second level video frame quadrants. A node is a leaf when its corresponding video frame quadrants conform to the split criteria; otherwise, the nodes will be called non-terminal or internal (see Fig. 2). We used the Z-order function, following the NW, NE, SW, SE directions.

First numeral 0 identifies the initial quadrant representing the whole video sequence and the tree root. Numerals $1, \ldots, T$, following their