Signal Processing 110 (2015) 27–36

Contents lists available at ScienceDirect

Signal Processing

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

Dynamic spatio-temporal modeling for example-based human silhouette recovery

Xue Zhou [a], Xi Li [b,*]

[a] School of Automation Engineering, University of Electronic Science and Technology of China, China
[b] College of Computer Science, Zhejiang University, China

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form
30 July 2014
Accepted 12 August 2014
Available online 22 August 2014

Keywords:
Human silhouette recovery
Spatio-temporal modeling
Dynamic time warping
Data alignment
Nonnegative matrix factorization

ABSTRACT

In this paper, we pose human silhouette recovery as a problem of robust spatio-temporal signal restoration, which aims to effectively recover the original human silhouette signals from noisy corruption or partial occlusion by investigating their intrinsic structural properties in both spatial and temporal dimensions. In this case, the underlying temporal correlations among adjacent silhouette frames are discovered by solving an adaptive time-series data alignment optimization problem using dynamic time warping (DTW). Furthermore, we build a part-based shape model to capture the spatial structural information on human silhouettes by sparseness constrained nonnegative matrix factorization (NMF)-based local feature learning, which is capable of well modeling the shape variation properties of human silhouettes. Experimental results on several challenging datasets demonstrate the effectiveness of our method.

1. Introduction

Human pose recovery is an important issue for many computer vision applications, including video indexing, surveillance, automotive safety and behavior analysis, as well as many other human computer interaction applications [1–3]. Human pose recovery is also a challenging problem that involves estimating many degrees of freedom and has many different taxonomies.

From visual input, rather than using the original feature, image is often described in terms of edges, color, texture, motion or silhouettes as most of the body pose information remaining in its silhouette and contour (silhouette outline). Thus, in this paper, we mainly focus on human silhouette recovery and its scope is limited to effectively restore the original silhouette signals from exemplar database while the original signals are in the presence of complicated factors (e.g. noise or partial occlusion). A solution to robust silhouette signal restoration is to explore the underlying spatio-temporal structural information on human silhouette sequences. Therefore, we focus on mining the temporal correlations across silhouette frames and modeling the spatial shape properties of human silhouettes.

More specifically, the task of temporal correlation mining is accomplished by solving a time-series data matching optimization problem, which is associated with a dynamic time warping (DTW) process for two human silhouette sequences with unequal lengths. The optimal solution to the optimization problem is obtained by dynamic programming. Moreover, the similarity measure used in DTW largely depends on what kind of shape descriptor has been chosen. Different from the global shape statistical feature, part-based methods are capable of modeling local shape variations which is much more robust to noise corruption and local occlusion. Especially for nonnegative matrix factorization (NMF) method, only additive, not subtractive combinations are allowed and

* Corresponding author.
E-mail addresses: zhouxue@uestc.edu.cn (X. Zhou), xilizju@zju.edu.cn (X. Li).

http://dx.doi.org/10.1016/j.sigpro.2014.08.019
0165-1684/

some clutter can be well suppressed or downweighted. Thus, in our method, the task of spatial shape modeling is performed by constructing a part-based shape model that captures the intrinsic subspace information on human silhouettes. The problem of part-based shape model construction is converted to that of NMF-based local feature learning, which tries to learn a set of part-based shape bases for characterizing local shape variations.

Therefore, the contributions of this work lie in the following three aspects: (1) We introduce the spatio-temporal modeling into the process of human silhouette recovery, that is, utilizing the intrinsic spatio-temporal information leads to the significant improvement of silhouette signal restoration. (2) We formulate the problem of temporal correlation mining as a dynamic programming based time-series data matching problem, which enforces the temporal smoothness constraints on human silhouette recovery. (3) We propose a sparseness constrained NMF-based linear shape model that is capable of well utilizing the local shape information to reduce the influence of partial occlusion or noisy corruption.

2. Related work

Following the survey paper [3], the human pose recovery can be mainly classified into two classes: model-based and model-free approaches. Model-based methods rely on a prior kinematic or shape model in either 2D or 3D [4,5]. The pose estimation process consists of modeling and estimation. If no explicit human body model is available, a learning-based or an example-based method (belonging to model free methods) is adopted to estimate the direct relation between image observation and pose. In learning-based methods, maximum a posteriori estimate [6] and regression-based [7,8] methods are widely used to learn a function from image space to pose space using training data. Recently, some manifold learning (ML)-based methods try to learn a low-dimensional latent space to preserve the geometric structure of the training data, including graph learning-based [9,10,14] and hypergraph learning-based methods [11–13]. Avoid learning this mapping, example-based methods store a dataset of exemplars, together with their corresponding pose descriptions [15–18]. The recovered pose is estimated by searching for the most similar exemplars and performing an interpolation among their corresponding candidate poses.

Temporal correlation mining is necessary to be considered when finding the nearest neighbor in example-based pose recovery methods. Toyama and Blake [15] incorporate exemplars in a probabilistic temporal framework. The Markov chain is used to achieve temporal coherence in Howe's work [16]. Zhou et al. [18] construct a subsequence by concatenating the test sample with its previous adjacent frames, and transform the recovery problem to a "shortest path" searching problem. Other than considering the test frame alone, considering the subsequence as a whole unit and matching two time series data are much more robust to noise disturbance.

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two time-dependent sequences under certain restrictions [19]. Originating from speech processing, nowadays DTW has been successfully applied in fields such as shape retrieval, matching and classification [20–23]. In many real applications, for a shorter query fragment, the demand of finding the most similar subsequence within the longer database is much more common. Instead of aligning the two sequences globally, a variant of classic DTW called the subsequence DTW has emerged [19].

Due to embedding lots of shape information, silhouettes or contours (silhouette outlines) are often used for describing poses. Traditionally, silhouettes are encoded using central moments [24] or Hu moments [25]. Contours can be encoded by shape contexts [26] or level sets [27]. For a given corrupted or noise disturbed input, a similarity search based on the above features may be sensitive to the noise disturbance. In order to improve the robustness, sparse shape representation is getting more and more popular [28,29]. The one reason is because given a large enough training dataset, an input can be approximately by a sparse linear combination of training shapes and the local detail information which is not statistically significant can be recovered with such a setting. Moreover, the input may contain gross errors, but such errors are often very sparse. Based on the above analysis, the local part-based sparse shape representation is suitable for human silhouette recovery.

3. Proposed method

3.1. Problem definition

Given a human silhouette training sequence $Y=(y_1, y_2, \ldots, y_M)$ and any test sample x_t which may be seriously corrupted from a test sequence $X=(x_1, x_2, \ldots, x_t, \ldots, x_N)$, our goal is to reconstruct x_t using its r-nearest exemplars from training dataset:

$$\hat{x_t} \approx \sum_{i=1}^{r} w_i y_{ne_i} \tag{1}$$

where ne_i is the index of the ith nearest exemplars from training dataset Y, w_i is the weight and the reconstructed sample is denoted by $\hat{x_t}$.

Different from the traditional methods which treat test samples independently, we explore the underlying spatio-temporal structural information embedded in the time series data and convert the recovery problem to a subsequence matching problem. Specifically, we concatenate x_t with its previous $L-1$ samples to form a short query fragment $Q=(x_{t+1-L}, \ldots, x_t) \subseteq X$ of length L such that $L \ll M$. Then the objective is to find a subsequence (indexed from a^* to b^* as shown in Fig. 1) within the longer database sequence Y that is most similar to the query fragment Q. Thus by introducing the temporal consistency among consecutive frames, the reliable r-nearest exemplars can be found. The above subsequence matching problem can just be solved by subsequence DTW algorithm. Fig. 1 is the illustration of the subsequence DTW. In the following, we describe how to utilize the spatio-temporal data mining for human silhouette recovery. Some important notations with their descriptions are presented in Table 1.