Image and Vision Computing 31 (2013) 291–310

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Editor's choice article

Homeomorphic Manifold Analysis (HMA): Generalized separation of style and content on manifolds ☆

Ahmed Elgammal [a,*], Chan-Su Lee [b]

[a] Department of Computer Science, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854, USA
[b] Department of Electronic Engineering, Yeungnam University, 214-1 Dae-dong, Gyeongsan-si, Gyeongbuk-do 712-749, Republic of Korea

ARTICLE INFO

Article history:
Received 9 April 2012
Received in revised form 28 September 2012
Accepted 24 December 2012

Keywords:
Style and content
Manifold embedding
Kernel methods
Human Motion Analysis
Gait analysis
Facial expression analysis

ABSTRACT

The problem of separation of style and content is an essential element of visual perception, and is a fundamental mystery of perception. This problem appears extensively in different computer vision applications. The problem we address in this paper is the separation of style and content when the content lies on a low-dimensional nonlinear manifold representing a dynamic object. We show that such a setting appears in many human motion analysis problems. We introduce a framework for learning parameterization of style and content in such settings. Given a set of topologically equivalent manifolds, the Homeomorphic Manifold Analysis (HMA) framework models the variation in their geometries in the space of functions that maps between a topologically-equivalent common representation and each of them. The framework is based on decomposing the style parameters in the space of nonlinear functions that map between a unified embedded representation of the content manifold and style-dependent visual observations. We show the application of the framework in synthesis, recognition, and tracking of certain human motions that follow this setting, such as gait and facial expressions.

1. Introduction

The problem of separation of style and content is an essential element of visual perception and is a fundamental mystery of perception [1,2]. For example, we are able to recognize faces and actions under wide variability in the visual stimuli. While the role of manifold representations in perception is still unclear, it is clear that images of the same object lie on a low-dimensional manifold in the visual space defined by the retinal array. On the other hand, neurophysiologists have found that neural population firing is typically a function of a small number of variables, which implies that population activities also lie on low-dimensional manifolds [1].

In this paper we consider the visual manifolds of biological motion. Despite the high dimensionality of the configuration space, many human motions intrinsically lie on low-dimensional manifolds. This is true for the kinematics of the body, as well as for the observed motion through image sequences. Let us consider the observed motion. For example, the silhouette (occluding contour) of a human walking or performing a gesture is an example of a dynamic shape, where the shape deforms over time based on the action being performed. These deformations are restricted by the physical body and the temporal constraints posed by the action being performed. Given the spatial and the temporal constraints, these silhouettes, as points in a high-dimensional visual input space, are expected to lie on a low-dimensional manifold. Intuitively, the gait is a one-dimensional manifold that is embedded in a high-dimensional visual space. This was also shown in [3,4]. Such a manifold can be twisted and even self-intersect in the high-dimensional visual space. Similarly, the appearance of a face performing expressions is an example of a dynamic appearance that lies on a low-dimensional manifold in the visual input space.

Although the intrinsic body configuration manifold might be very low in dimensionality, the resulting visual manifold (in terms of shape and/or appearance) is challenging to model, given the various aspects that affect the appearance. Examples of such aspects include the body type (slim, big, tall etc.) of the person performing the motion, clothing, viewpoint, and illumination. Such variability makes the task of learning a visual manifold very challenging, because we are dealing with data points that lie on multiple manifolds at the same time: body configuration manifold, viewpoint manifold, body shape manifold, illumination manifold, etc.

The main contribution of this paper is a novel computational framework for learning a decomposable generative model that explicitly factorizes the intrinsic body configuration (content), as a function of time, from the appearance (style) factors. The framework we present in this paper is based on decomposing the style parameters in the

☆ This paper has been recommended for acceptance by Maja Pantic. Editor's Choice Articles are invited and handled by a select rotating 12 member Editorial Board committee.

* Corresponding author. Tel.: +1 732 445 2001x0021; fax: +1 732 445 0537.
E-mail addresses: elgammal@cs.rutgers.edu (A. Elgammal), chansu@ynu.ac.kr (C.-S. Lee).

0262-8856/$ – see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.imavis.2012.12.003

space of nonlinear functions that maps between a unified representation of the content manifold and style-dependent observations. Given a set of topologically equivalent manifolds, the Homeomorphic Manifold Analysis (HMA) framework models the variation in their geometries in the space of functions that maps between a topologically-equivalent common representation and each of them. The common representation of the content manifold can be learned from the data or can be enforced in a supervised way if the manifold topology is known. The main assumption here is that the visual manifold is homeomorphic to the unified content manifold representation, and that the mapping between that unified representation and the visual space can be parameterized by different style factors. We describe the motivations and contributions of the framework in more detail within the context of the state-of-the-art in Section 2.

The learned models support tasks such as synthesis and body configuration recovery, as well as the recovery of other aspects such as viewpoint, person parameters, etc. As direct and important applications of the introduced framework, we consider the cases of gait and facial expressions. We show an application of the framework to gait analysis, where the model can generate walking silhouettes for different people from different viewpoints. Given a single image or a sequence of images, we can use the model to solve for the body configuration, viewpoint, and person shape style parameters. We also show the application of the framework to facial expressions as an example of a dynamic appearance. In this case, we learn a generative model that generates different dynamic facial expressions for different people. The model can successfully be used to recognize expressions performed by different people who are not used in model training, as well as identifying the person performing the expression.

The paper organization is as follows, Section 2 discusses the relation between the proposed framework and the state-of-the-art. Section 3 summarizes the framework and its applications. Section 4 describes different ways to obtain a unified content manifold representation, in both unsupervised and supervised ways. Section 5 describes the details for learning the factorized model. Section 6 describes algorithms for solving for multiple factors. Section 7 shows experimental results and examples of applying the model for dynamic shape and appearance manifolds, for the analysis of gait and facial expressions.

2. Relation to state-of-the-art

This section puts the contributions of the paper in the context of the state-of-the-art in related areas.

2.1. Factorized models: Linear, bilinear, and multi-linear models

Linear models, such as PCA [5], have been widely used in appearance modeling to discover subspaces for appearance variations. For example, PCA has been used extensively for face recognition, such as [6–9], and to model the appearance and view manifolds for 3D object recognition, as in [10]. Subspace analysis can be further extended to decompose multiple orthogonal factors using bilinear models and multilinear tensor analysis [11,12]. The pioneering work of Tenenbaum and Freeman [11] formulated the separation of style and content using a bilinear model framework [13]. In that work, a bilinear model was used to decompose face appearance into two factors: head pose and different people as style and content interchangeably. They presented a computational framework for model fitting using SVD. Bilinear models have been used earlier in other contexts [13,14]. A bilinear model is a special case of a more general multilinear model. In [12], multilinear tensor analysis was used to decompose face images into orthogonal factors controlling the appearance of the face including geometry (people), expressions, head pose, and illumination using High Order Singular Value Decomposition (HOSVD) [15]. Tensor representation of image data was used in [16] for video compression, and in [17] for motion analysis and synthesis. N-mode analysis of higher-order tensors was originally proposed and developed in [13,18,19] and others. The applications of bilinear and multilinear models to decompose variations into orthogonal factors, as in [11,12], are mainly for static image ensembles.

The question we address in this paper is how to separate the style and content on a manifold representing a dynamic object. Why don't we just use a bilinear model to decompose the style and content in this case, where certain body poses can be denoted as content and different people as style? The answer is that in the case of dynamic (e.g. articulated) objects, the resulting visual manifold is nonlinear. This can be illustrated by considering the example walking cycle in Fig. 2. In this case, the shape temporally undergoes deformations and self-occlusion, which results in a nonlinear manifold. The two shapes in the middle of the two rows correspond to the farthest points in the walking cycle kinematically, which are supposedly the farthest points on the manifold, in terms of the distance along the manifold. In the Euclidean visual input space, these two points are very close to each other, as can be noticed from the distance plot on the right of Fig. 2. Because of such nonlinearity, PCA, bilinear, and multilinear models will not be capable of discovering the underlying manifold and decomposing the orthogonal factors. Linear models will not be able to interpolate intermediate poses and/or intermediate styles.

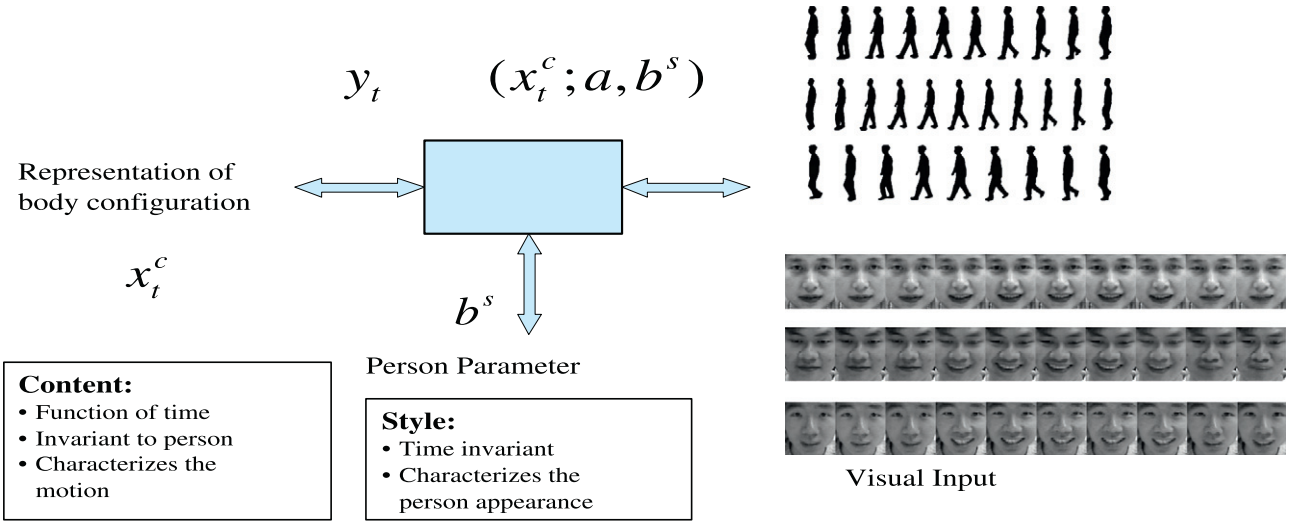

Fig. 1. Style and content factors. Content: gait motion or facial expression. Style: different silhouette shapes or face appearance.