Pattern Recognition Letters 50 (2014) 130–138

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Human activity recognition by separating style and content

Muhammad Shahzad Cheema [a,*], Abdalrahman Eweiwi [a], Christian Bauckhage [a,b]

[a] Bonn Aachen International Center for IT, University of Bonn, Dahlmannstrasse 2, 53113, Bonn, Germany
[b] Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augsitin, Germany

ARTICLE INFO

Article history:
Available online 4 October 2013

Communicated by Dmitry Goldgof

Keywords:
Action recognition
Bilinear models
Kinect depth images
Motion history volumes
Motion capture
Expectation maximization

ABSTRACT

Studies in psychophysics suggest that people tend to perform different actions in their own style. This article deals with the problem of recognizing human actions and the underlying execution styles (actors) in videos. We present a hierarchical approach that is based on conventional action recognition and asymmetrical bilinear modeling. In particular, we employ bilinear factorization on the tensorial representation of the action videos to characterize styles of performing different actions. Our approach is solely based on the dynamics of the underlying activity. The model is evaluated on the IXMAS and the Berkeley-MHAD data sets using different modalities based on optical motion capture, Kinect depth videos, and 3D motion history volumes. In each case high recognition accuracy is achieved in comparison to the symmetric bilinear modeling and the Nearest Neighbor classification.

1. Introduction

Recognizing human actions in videos has become a rapidly growing area of research. Most existing research has focused exactly on this very aspect. However, people tend to execute different actions and activities such as walking, kicking and cooking in their own personal manner. Studies in psychophysics and biomechanics have shown that individuals build specific internal models for different movements and they can be recognized solely from their motion dynamics (Cutting and Kozlowski, 1977; Thoroughhman and Shadmehr, 1999). In the last decade, corresponding research on vision-based gait recognition (Wang et al., 2010) has shown great success. This motivates us to investigate the generalized multi-label classification scenario for human activity recognition. In particular, we are interested to determine *if it is possible to recognize both the actions and the actors in videos?* This problem is related to well-known problem of *separating style from content* in areas such as handwriting or face recognition. The research presented in this article builds on our previous work (Cheema et al., 2012). To the best of our knowledge, our approach is the first at such an attempt in the context of human action recognition.

Identifying style and content has been of great interest for the recognition of handwriting, speech, and faces since the idea was pioneered by Tanenbaum and Freeman (2000). There, the authors applied bilinear modeling and showed promising results on classical problems such as handwritten character, face, or pose recognition. Chaung and Bregler (2005) used bilinear factorization to separate emotional styles from speech contents in order to create expressive facial animations. Shin et al. (2008) proposed an efficient approach to "illumination-robust" face recognition, based on symmetric bilinear modeling, by separating an identity factor and an illumination factor.

Despite a great deal of research on human action recognition (Laptev et al., 2008; Marszalek et al., 2009; Krausz and Bauckhage, 2010; Wang et al., 2011; Reddy and Shah, 2012), hardly any efforts have yet been made to separate style from content. Most of the existing work in this direction deals with person identification for a single action (Elgammal and Lee, 2004; Cuzzolin, 2006; Perera et al., 2009; Lee and Elgammal, 2004). Elgammal and Lee (2004) applied a non-linear model for separating poses from walking patterns of individuals; Cuzzolin (2006) used bilinear separation models to different gait gestures. The approach presented in Yam et al. (2002) was the first to consider styles of running in recognizing individuals. Perera et al. (2009) employed multifactor tensor decomposition to identify different styles of the dancing action using motion capture data. Recently, Iosifidis et al. (2011) trained person specific activity classifiers to improve recognition of different human actions. The issue of varying styles for human activities has been discussed by Taralova et al. (2011), which presents a source constrained clustering approach to accommodate different sources (e.g. actors). However, their focus is on clustering from *known* sources and not on identifying the sources. Our previous work (Cheema et al., 2012) establishes applicability of bilinear modeling towards human activity recognition in a constrained environment using motion history volumes.

The approach presented in this article treats observed actions as resulting from a generative process with two factors, namely actor

* Corresponding author. Tel.: +49 228 2699131.
E-mail addresses: cheema@bit.uni-bonn.de, m.shahzad.cheema@gmail.com (M.S. Cheema), eweiwi@bit.uni-bonn.de (A. Eweiwi), christian.bauckhage@iais.fraunhofer.de (C. Bauckhage).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.024

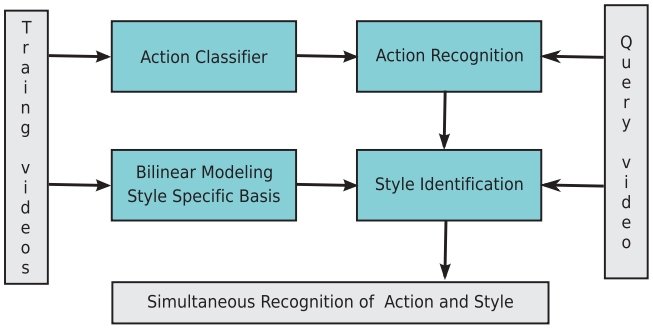

Fig. 1. Block diagram of our approach.

(style) and action (content). We use bilinear factorization to model underlying phenomena since bilinear models immediately lend themselves towards two-factor classification and since they can be efficiently determined through singular value decompositions (SVDs). Conventional symmetric bilinear models assume independence between content and style factors (e.g. face and illumination). This is not the case in motion based action recognition since both actions and the execution styles are based on the variation of the same cue, i.e. motion. Due to challenges posed by high articulation of human bodies, conventional symmetric bilinear model would not suffice to separate content and style in human activity videos. We therefore use a two-step approach to classify a given test video (query video). In the first step, we apply a classical action classification to predict underlying action of the query video. In the second step, we use this prediction to generate a style-specific basis for the query video using an asymmetric bilinear model. Finally, we compare this basis with the style-specific basis learned from training data in order to identify the most likely style. Fig. 1 gives an overview of the proposed approach. For experimental evaluation, we consider two multi-actor multi-action data sets namely Inria Xmas Motion Acquisition Sequences (IXMAS) (Weinland et al., 2006) and Berkeley Multimodal Human Action Database (MHAD) (Ofli et al., 2013). We show that, compared to naive nearest neighbor classification and symmetric bilinear modeling, the proposed hierarchical model significantly improves results for different motion cues. Consequently our approach extends motion-based person identification to multiple common actions and show that the identification is not limited to walking or running actions.

The remainder of the article is organized as follows: Section 2 reviews the basics of bilinear models. Section 3 presents how we deal with action recognition by using nearest neighbor classifiers and asymmetric bilinear models. Section 4 reports details on our benchmark data, experiments, and results. Finally, Section 5 concludes the work.

2. Bilinear models

In this section, we review basic concepts of bilinear models for separating style from content; our terminology is similar to the one used by Tanenbaum and Freeman (2000). A bilinear model is a generative model where each K dimensional observation $\mathbf{y}$ in a style $s \in [1,2,\ldots,S]$ and content class $c \in [1,2,\ldots,C]$ is given in the form:

$$y_k^{sc} = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ijk} a_i^s b_j^c, \quad k \in [1,2,\ldots,K] \tag{1}$$

where $\mathbf{a}^s$ and $\mathbf{b}^c$ are I and J dimensional coefficient vectors representing style s and content c and the entries w_{ijk} govern interaction between the two underlying factors. Let $\mathbf{W}_k$ represent kth matrix of dimension $I \times J$ then Eq. (1) becomes:

$$y_k^{sc} = \mathbf{a}^{sT}\mathbf{W}_k\mathbf{b}^c \tag{2}$$

The matrices $\mathbf{W}_k$ define bilinear mapping from content and style space to the K dimensional observation space. The model in Eq. (1) and Eq. (2) is called the *symmetric bilinear model*.

While the symmetric model assumes independence of the interaction terms w_{ijk} w.r.t style and content classes, the *asymmetric bilinear model* lets these terms vary with one of the factors (by convention with style) and thus allows for more flexibility. For instance, with a style-specific basis $\mathbf{a}_{jk}^s = \sum_i a_i^s w_{ijk}$, Eq. (1) becomes:

$$y_k^{sc} = \sum_{j=1}^{J} \mathbf{a}_{jk}^s b_j^c \tag{3}$$

Equivalently in matrix notation we write

$$\mathbf{y}^{sc} = \mathbf{A}^s\mathbf{b}^c \tag{4}$$

such that $\mathbf{A}^s$ denotes $K \times J$ matrix with entries $\mathbf{a}_{jk}^s$. Here, $\mathbf{A}^s$ represents a style-specific map from the content space to observation space.

2.1. Training an asymmetric bilinear model

Let $\mathbf{y}(t)$ denote tth training sample ($t = 1,\ldots,T$) and let $\boldsymbol{\chi}_{sc}(t)$ be a characteristic function such that $\boldsymbol{\chi}_{sc}(t) = 1$ if $\mathbf{y}(t)$ has style s and content c and 0 otherwise. Then sum of squared errors $\mathbf{E}$ for the asymmetric model over all training data is given by

$$\mathbf{E} = \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{c=1}^{C} \boldsymbol{\chi}_{sc}(t)\,\|y(t) - \mathbf{A}^s\mathbf{b}^c\|^2. \tag{5}$$

Fitting an asymmetric model aims at finding solutions for $\mathbf{A}^s$ and $\mathbf{b}^c$ that minimize $\mathbf{E}$. If a given sample of training data consists of nearly equal numbers of observations for each style and content (as in case of this article), a closed form procedure can be adopted from using SVD.

Let $\overline{\mathbf{y}}^{sc}$ denotes the *mean* of all observations in style s and content c, the training set can be thought of as a 3rd order tensor $\overline{\mathbf{Y}}_{S\times K\times C}$. For making efficient use of matrix algebra, $\overline{\mathbf{Y}}$ is represented as a stacked matrix with dimensions $(SK) \times C$ such that each of C columns contains S parts of $K \times 1$ vectors. Further, the *vector transpose VT* operation is employed to determine a $(BK) \times A$ matrix transpose of an $(AK) \times B$ stacked matrix, See details in Tanenbaum and Freeman (2000).

For training data in stacked matrix form, the asymmetric model can then be expressed as $\overline{\mathbf{Y}} = \mathbf{AB}$, such that $\mathbf{A} = \left[\mathbf{A}^1 \ldots \mathbf{A}^S\right]'$ is a $(SK) \times J$ matrix of style-specific basis and $\mathbf{B} = \left[\mathbf{b}^1 \ldots \mathbf{b}^C\right]$ is a $J \times C$ is matrix of content parameters. A least squares optimal solution is obtained by computing the SVD of $\overline{\mathbf{Y}}$ such that $\overline{\mathbf{Y}} = \mathbf{U\Sigma V}^T$. The style-specific basis matrix $\mathbf{A}$ is obtained from the first J columns of US and the content parameter matrix $\mathbf{B}$ is defined by the first J rows of V^T.

2.2. Training a symmetric bilinear model

The sum of squared error for the symmetric model in Eq. (2) is

$$\mathbf{E} = \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{k=1}^{K} \boldsymbol{\chi}_{sc}(t)\,\|y_k(t) - \mathbf{a}^{sT}\mathbf{W}_k\mathbf{b}^c\|^2. \tag{6}$$

To solve this optimization problem, asymmetric modeling through SVD is iterated by alternatively switching the roles of content and style and an expectation maximization (EM) approach is used to