Neurocomputing 205 (2016) 507–518

Contents lists available at ScienceDirect

Neurocomputing

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

A semantic RBM-based model for image set classification ☆

S. Elaiwat [a,*], M. Bennamoun [a], F. Boussaid [b]

[a] School of Computer Science and Software Engineering, University of Western Australia, Australia
[b] School of Electrical, Electronic and Computer Engineering, University of Western Australia, Australia

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form
28 September 2015
Accepted 3 May 2016
Communicated by Yongzhen Huang
Available online 12 May 2016

Keywords:
Face recognition
Image set classification
Restricted Boltzmann Machine
Temporal features

ABSTRACT

Most existing image set classification methods use either appearance variations or temporal information to represent semantic knowledge (relationships) and subject appearance. Such methods usually rely on a predetermined surface structures that the image sets could lie on, and/or are highly influenced by the temporal correlations between images within the image sets. In contrast, this paper introduces a novel RBM-based model which is capable of combining both, appearance variations and temporal information within image sets, to provide an automated and robust representation of semantic knowledge and subject appearance even with small image sets with weak temporal correlations. The structure of the proposed model involves two hidden sets which are used to encode different feature types. The first hidden set is used to represent the dominant appearances (facial features) from appearance variations, while the second set is used to represent the temporal information between different appearances. An extension of the standard Constructive Divergence algorithm is proposed to learn the proposed model encoding two different feature types simultaneously, while isolating them from each other. The proposed model was evaluated for the task of face recognition, using two datasets, namely UCSD/Honda and YouTube Celebrities. The results show superior performance compared to state-of-the-art methods.

1. Introduction

In traditional face recognition approaches, referred hereafter to as single image-based classification approaches, a single probe image is used to define the identity of the subject while multiple images are used to build the gallery set (training data) for each subject [1–6]. This approach has severe limitations, notably in uncontrolled environment, given the large possible variations of the subject's appearance. These variations could stem from changes in lighting conditions, camera pose, facial expressions and/or non-rigid deformations. To address these limitations, a number of works have investigated the problem of face recognition from multiple-images, also known as image set classification problem [7–14]. Compared to single image based approaches, classification from an image set exploits a wide range of facial appearance variations within each set. Thus, image set classification has the potential to achieve a robust performance in practical applications [15–17]. However, the image set classification approach introduces a new challenge which consists in defining an efficient image set model which is capable of exploiting the semantic knowledge (relationships) between images within a set [11]. Tackling this challenge using traditional classification methods (e.g, K-Nearset Neighbor and SVM) is not feasible because these techniques can only deal with the classification from single images.

To address this challenge, existing image set classification methods rely either on the appearance variations or the temporal information within images in the sets. Some of the recent proposed appearance variation based methods model the variability in the individual's appearance within an image set, using predetermined surface structures such as linear/affine subspaces [18,19], non-linear manifolds [8,9], and/or a mixture of subspaces [20,21]. On the other hand, the temporal information based methods, which are widely used for video classification [22], exploit the temporal dynamic information between a sequence of images/frames. Examples of such methods include the standard generative models (e.g, Restricted Boltzmann Machines (RBMs) or Hidden Markov Model (HMMs) [23,24]), which have been used to encode the temporal information through latent variables, used to estimate the posterior distribution of a given image set. Recently, extensions of the standard generative models such as Gated-RBM have been introduced to learn the temporal information between images, while defining the conditional distribution of the current image (current frame) given previous time slice input images (previous frames). Defining such a conditional distribution has shown to produce a higher ability to encode the relationships between images compared to the standard distributions.

☆This research is supported by the Australian Research Council grant DP110102166.

* Corresponding author.
E-mail address: Elaiws01@student.uwa.edu.au (S. Elaiwat).

http://dx.doi.org/10.1016/j.neucom.2016.05.013
0925-2312/

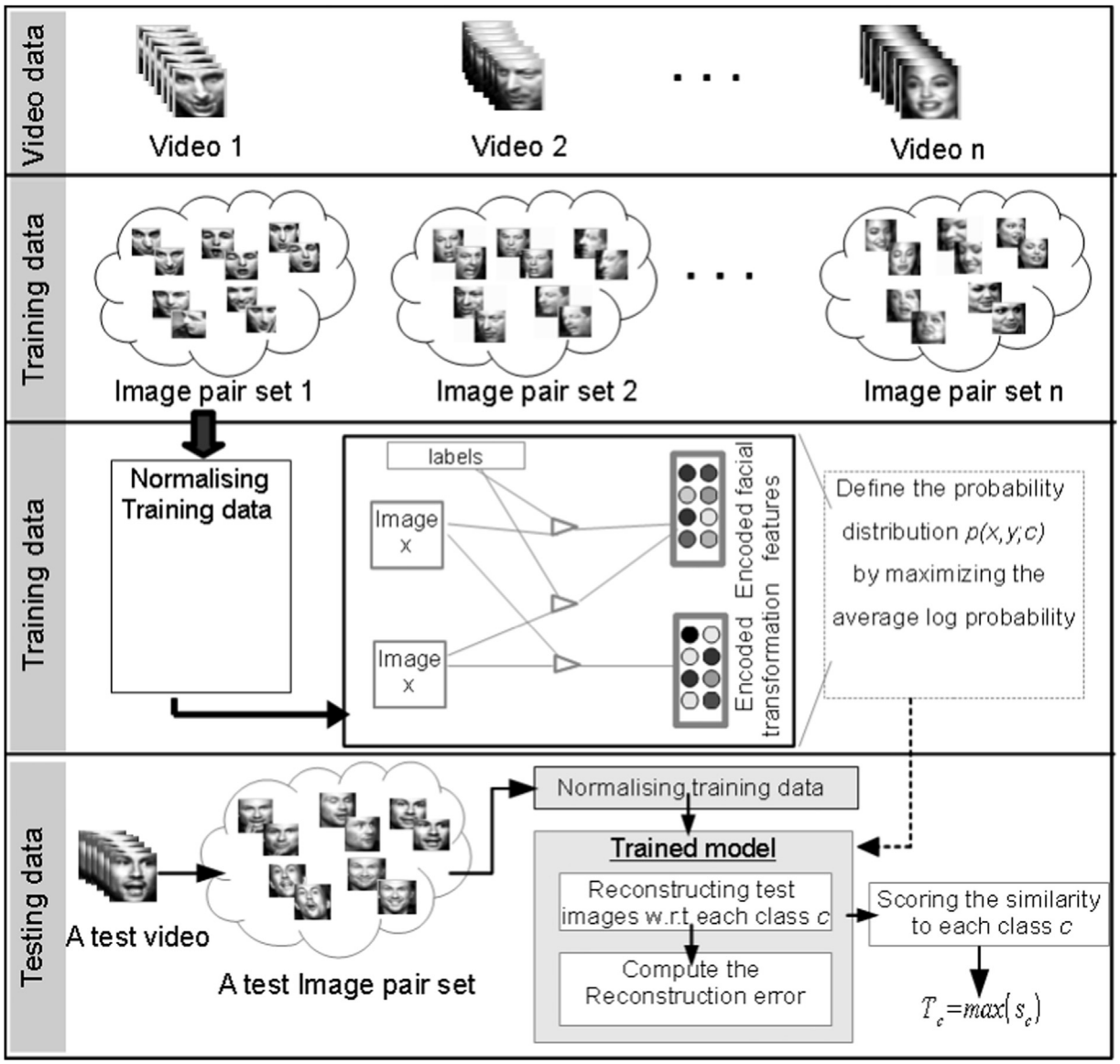

Fig. 1. Block diagram of the proposed model.

In this paper, we present a novel image set classification method for the problem of face recognition. The proposed model represents the semantic knowledge and the subject appearance by encoding both the facial features (dominant appearances) and the temporal information within sets of image pairs. This model exploits appearance variations to encode facial features (dominant appearances) while simultaneously encoding the temporal information between different appearances. This method provides therefore a robust representation of image sets even under challenging conditions (e.g., low resolution, occlusion and/or high pose variations). In addition, the proposed model is free from any prior assumption in terms of selecting the most appropriate surface structures (e.g, non-linear manifolds) to represent image sets. Rather, suitable nonlinear structures that are capable of efficiently representing both facial features and temporal information are defined automatically during the training phase. A block diagram of our proposed model is depicted in Fig. 1.

The main contribution of this work can be summarized as follows:

- A novel RBM-based model is proposed to learn the semantic knowledge and subject appearance within image sets by encoding both facial features (dominant appearance) and temporal information (relationships between images), simultaneously. This differs from other works, in that our model learns features using two different sets of latent variables (hidden units) associated with three sub-models. The first hidden set is dedicated to encode the facial features through two sub-models which define the conditional distribution of the input images (image pairs) given label units. These label units play an important role in learning facial features w.r.t. the identity of the subjects. The other hidden set is dedicated to encode the temporal information through a single sub-model, which defines the joint distribution between input images (image pairs). This method exploits the combination of encoded features (facial features and temporal information) to provide a more robust classification even in the presence of a weak relationship between training and testing data.
- An extension of the standard Contrastive Divergence algorithm (CD) is proposed to learn two different feature types (facial features and temporal information), simultaneously, and isolating them from each other. Compared to the standard CD algorithm, which only involves a single hidden set and a single input set, our model involves two input sets, two hidden sets