Image and Vision Computing 52 (2016) 125–140

Contents lists available at ScienceDirect

Image and Vision Computing

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

Improving facial analysis and performance driven animation through disentangling identity and expression☆

David Rim[a,1], Sina Honari[b,1], Md Kamrul Hasan[a], Christopher J. Pal[a,*]

[a]Département de génie informatique et génie logiciel École Polytechnique Montréal, Montréal Québec H3T 1J4, Canada
[b]Département d'informatique et de recherche opérationnelle Université de Montréal, Montréal, Québec H3C 3J7, Canada

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 24 March 2016
Accepted 21 April 2016
Available online 31 May 2016

Keywords:
Factorization techniques
Emotion recognition
Graphical models
Performance driven animation
Facial expression analysis

ABSTRACT

We present techniques for improving performance driven facial animation, emotion recognition, and facial key-point or landmark prediction using learned identity invariant representations. Established approaches to these problems can work well if sufficient examples and labels for a particular identity are available and factors of variation are highly controlled. However, labeled examples of facial expressions, emotions and key-points for new individuals are difficult and costly to obtain. In this paper we improve the ability of techniques to generalize to new and unseen individuals by explicitly modeling previously seen variations related to identity and expression. We use a weakly-supervised approach in which identity labels are used to learn the different factors of variation linked to identity separately from factors related to expression. We show how probabilistic modeling of these sources of variation allows one to learn identity-invariant representations for expressions which can then be used to identity-normalize various procedures for facial expression analysis and animation control. We also show how to extend the widely used techniques of active appearance models and constrained local models through replacing the underlying point distribution models which are typically constructed using principal component analysis with identity–expression factorized representations. We present a wide variety of experiments in which we consistently improve performance on emotion recognition, markerless performance-driven facial animation and facial key-point tracking.

1. Introduction

One of the primary sources of variation in facial images is identity. Although this is an obvious statement, many approaches to vision tasks other than facial recognition do not directly account for the interaction between identity-related variation and other sources. However, many facial image datasets are subdivided by subject identity and this provides additional information that is often unused. This paper deals with the natural question of how to effectively use identity information in order to improve tasks other than identity recognition. In particular, our primary motivations, applications of interest and goals are to develop methods for facial expression analysis and performance driven facial animation that are less identity dependant.

Recently there has been work on facial recognition in which identity is separated from other sources of variation in 2D image data in a fully probabilistic way [37]. In this model, the factors are assumed to be additive and independent. This procedure can be interpreted as a probabilistic version of Canonical Correlation Analysis (CCA) presented by Bach [2], or as a standard factor analysis with a particular structure in the factors. In this paper, we investigate and extend the use of this probabilistic approach to separate sources of variation, but unlike prior work which focuses on inferences about identity, we focus on facial expression analysis and facial animation tasks. This includes performance-driven animation, emotion recognition, and key-point tracking. Our goal is to use learned representations so as to create automated techniques for expression analysis and animation that better generalize across identities. We show here how disentangling factors of variation related to identity can indeed yield improved results. In many cases we show how to use the learned representations as input to discriminative classification methods.

In our experimental work, we apply this learning technique to a wide variety of different input types including: raw pixels, key-points and the pixels of warped images. We evaluate our approach by predicting: standard emotion labels, facial action units, and 'bone'

☆ This paper has been recommended for acceptance by Tim Cootes.
* Corresponding author.
E-mail addresses: daverim@gmail.com (D. Rim), sina.honari@umontreal.ca (S. Honari), md-kamrul.hasan@polymtl.ca (M. Hasan), christopher.pal@polymtl.ca (C. Pal).
[1] Joint first authors.

http://dx.doi.org/10.1016/j.imavis.2016.04.017
0262-8856/

positions or animation sliders which are widely used in computer animation. We go on to show how it is indeed possible to improve the facial keypoint prediction performance of active appearance models (AAMs) on unseen identities as well as increase the performance of constrained local models (CLMs) through our identity–expression factorization extensions to these widely used techniques. Our evaluation tasks and a comparison of the types of data being used as input for each experiment are summarized in Table 1.

The rest of this manuscript is structured as follows: In Section 2, we discuss the facial expression analysis, performance driven animation and keypoint tracking applications that serve as the ultimate goals of our work in more detail and review relevant previous work. The various methods we present here build in particular on the work of Prince et al. [37] in which a linear Gaussian probabilistic model was proposed to explicitly separate factors of variation due to identity versus expression. In Section 3, we present this model as a way to disentangle factors of variation arising from identity and expression variation. While Prince et al. used this type of identity–expression analysis to make inferences about identity, our work here focuses on how disentangling such factors can be used to make inferences about facial expression. Indeed, as discussed above, facial expression analysis and the applications of detailed facial expression analysis to computer animation is the motivating goal of our work here.

The first set of contributions of our work are presented in Section 3.2. These contributions consist of a wide variety of novel techniques for using learned identity–expression representations for common goals related to expression analysis and computer animation. We provide novel techniques and experiments in which we predict: emotion labels, facial action units, facial keypoints, and animation control points as summarized in Table 1. Of particular note is the fact that we propose a novel formulation for identity normalizing facial images which we use for facial action unit recognition, emotion recognition and animation control. We find that using this identity-normalized representation leads to improved results across this wide variety of the expression analysis tasks.

We also go on to extend the underlying identity and expression analysis technique in two important directions, providing two additional technical contributions. First, in Section 4 we show that identity–expression analysis can be used in place of the principal component analysis (PCA) technique that is widely used in active appearance models (AAMs). This procedure yields what we call an identity–expression factorized AAM, or IE-AAM. We present the modifications that are necessary to integrate this approach into the AAM framework of [34]. Our experiments show that IE-AAMs can increase the performance of PCA-AAMs dramatically when no training data is available for a given subject. We also found that the use of IE-AAMs eliminated the convergence errors observed with PCA-AAMs. Secondly, in Section 5 we show that constrained local models (CLMs) can also be reformulated, extended and improved through using an underlying identity–expression analysis model. Our reformulation also provides a novel energy function and minimization formulation for CLMs in general.

Table 1
Summary of experiments described in this paper, numbers of the section describing each experiment are given in parenthesis.

JAFFE	
Emotion recognition *Predicts:* Emotion labels *Using:* Images	Section 3.2.1
Animation control *Predicts:* Bone position parameters *Using:* Images	Section 3.2.3
Extended Cohn–Kanade	
Facial action unit (AU) detection *Predicts:* AU labels *Using:* Point locations Shape-normalized images Combined point locations and shape-normalized images	Section 3.2.2
Emotion recognition *Predicts:* Emotion labels *Using:* Point locations Shape-normalized images Combined point locations and shape-normalized images	Section 3.2.2
Key-point localization (IE-AAMs) *Predicts:* Key-point locations *Using:* Images	Section 4.3
CMU multi-PIE	
Key-point localization (IE-CLMs) *Predicts:* Key-point locations *Using:* Images	Section 5.4
Animation control studio data	
Animation control *Predicts:* Bone position parameters *Using:* Shape-normalized images	Section 3.2.4

2. Our applications of interest and relevant prior work

2.1. Performance-driven facial animation

Performance-driven animation is the task of controlling a facial animation via images of a performer. This is a common process in the entertainment industry. In many cases, this problem is generally handled by the placement of special markers on the performer [52]. However, here we are interested in developing markerless motion capture techniques in which we employ machine learning methods and minimize manual intervention.

Many marker-less facial expression analysis methods rely either on tracking points using optical flow [16] or fitting Active Appearance Models [30]. Morphable models in [6], [36], [40] have also been investigated. More recently, 3D data and reconstruction is used to fit directly to the performer [51], [54]. These methods, however, often require additional data, for example, multiview stereo [53], [4] or structured light [7], [51]. Sandbach et al. provide a thorough review of 3D expression recognition techniques in [38]. In the end, these methods usually work by providing dense correspondences which then require a re-targeting step.

However, a simpler and often used approach in industry does not rely on markers and simply uses the input video of a facial performance. The idea is to use a direct 2D to 3D mapping based on regressing image features [21] to 3D model parameters. This method works well but is insufficiently automatic. Each video is mapped to 3D model parameters, possibly with interpolation between frames. Key-point based representations often require both data and training time that compares unfavorably to this simpler approach given the re-targeting step. The primary benefit of key-point based representations appears to be a degree of natural identity-invariance. We provide experiments on performance driven animation, in which we predict bone positions using the well known Japanese Female Facial Expression (JAFFE) Database [33] as input as well as an experiment using professional helmet camera based video used for high quality, real world animations.

2.2. Emotion recognition

There is a large amount of overlap in the objectives for performance-driven facial animation and the objectives of detailed emotion recognition. Automatic emotion recognition has largely focused on the facial action coding system or FACS [15] as an auxiliary task [46]. In our work here we shall predict action unit (AU) values that have been annotated for the well known extended Cohn