Pattern Recognition 48 (2015) 759–771

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning from weakly labeled faces and video in the wild

David Rim, Md Kamrul Hasan, Fannie Puech, Christopher J. Pal*

Département de génie informatique et génie logiciel, École Polytechnique Montréal, Université de Montréal, Montréal, Québec, Canada

ARTICLE INFO

Article history:
Received 16 May 2013
Received in revised form
22 August 2014
Accepted 18 September 2014
Available online 30 September 2014

Keywords:
Semi-supervised learning
Face recognition
Graphical models

ABSTRACT

We present a novel method for face-recognition based on leveraging weak or noisily labeled data. We combine facial images from the Labeled Faces in the Wild (LFW) dataset with face images extracted from videos on YouTube and face images returned using a search engine. Our technique is based on a novel formulation for weakly supervised learning based on probabilistic graphical models using a margin-like property and a null category. As such, our formulation remains within a fully probabilistic framework. We use this technique to combine high accuracy human labeled data with noisily labeled data. We present a specific variation of our general approach using a model inspired by the relevance vector machine (RVM), a probabilistic alternative to support vector machines. In contrast to previous formulations of RVMs we show how the choice of an exponential hyperprior produces an approximation to the L_1 penalty. We present both experiments where we simulate noisy labels and experiments where we use image and video search results as noisily labeled data. Faces extracted from the resulting Youtube videos thus are likely, but not assured to contain examples of the person whose name was given as the query. We show how our probabilistic margin approach provides a robust way to combine labeled LFW data with this type of noisy search result. Our results indicate that recognition performance can indeed be increased consistently with weakly labeled data using our technique.

1. Introduction

Facial recognition research has begun to focus on the difficult task of recognition in unconstrained images, *i.e.* images of faces in commonly occurring conditions. These images, such as those usually produced by consumer digital photographers under conditions of varying illumination and pose, are often occluded (*e.g.* with glasses), and sometimes heavily compressed or degraded by motion blurring. Facial recognition in these conditions remains a difficult task, which is exacerbated when sufficient quantities of labeled data are not available.

However, finding sources of unlabeled facial images is not a challenging task. Several large web-based collections such as Flickr and Google Images provide public access to millions of static images. Video sites such as Youtube provide videos which can provide more information than static images alone.

Although the resulting data is unlabeled, useful information is often contained within. In this paper, we investigate the use of search queries as weak labels, which is visualized in Fig. 1. We show that unconstrained facial recognition in both static and video images can be significantly improved using this approach to unlabeled data.

1.1. Related work

Related work on facial recognition: Early face recognition systems [1–4] were based and tested on upright, frontal face images in datasets such as the AT&T ORL Database of Faces, or on constrained databases such as the FERET database [5]. These datasets were therefore amenable to factor analysis approaches such as principal component analysis (PCA) [6,1], Fisher's linear discriminant (FLD), [4], independent component analysis (ICA) [7], non-negative matrix factorization (NMF) [8], and probabilistic PCA (PPCA) [9], where the principal source of variation is assumed to be identity. Non-linear representations, such as Locality Preserving Projection (LPP), a linear graph embedding method [10,11] and Sparsity Preserving Projections (SPP) [12], used for locally linear manifold representations have been widely applied to face recognition. Similarity metrics, used commonly in Nearest Neighbor (NN) methods, have also shown good performance [13]. More recently, Wang et al. used subspaces based on Grassmann manifolds created by merging KD-tree leaf partitions based on distance metrics and classification rates in order to improve facial recognition [14]. However, most of these methods have been applied to controlled databases. Yang et al. describe similar face detection issues with complex background variation [15]. Zhang et al. provide an excellent review of the problems that arise due to uncontrolled pose variation in [16].

* Corresponding author.
E-mail address: christopher.pal@polymtl.ca (C.J. Pal).

http://dx.doi.org/10.1016/j.patcog.2014.09.016
0031-3203/

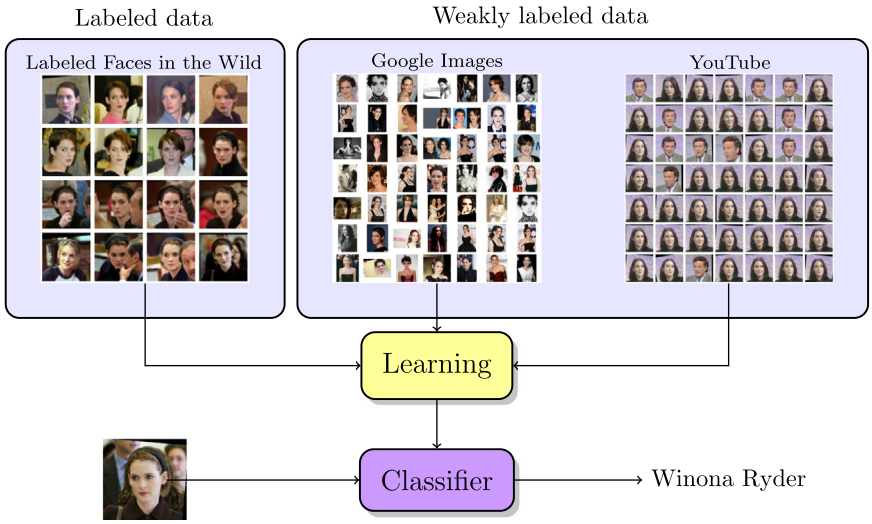

Fig. 1. High performance recognition requires a large number of labeled examples in the uncontrolled case. Large amounts of weakly labeled data are easily obtainable through the use of image and video search tools. Many of these examples are either irrelevant or not the identity in question. We learn a classifier that does not require manual labeling of the weakly labeled data by accounting for the weak label noise.

As a response to the growing need for less constrained databases as well as the more pragmatic reason of common evaluation, the Labeled Faces in the Wild Dataset (LFW) was created by Huang et al. [17] in order to provide a far more natural composition of face images. That is, additional sources of variation are also present, including image variation from pose, expression, illumination and background. However, an important element of *in the wild* face data sets is that the mean number of training images per subject is often limited. In the case of the LFW, the average number of labeled images is a little more than 2. As such, multiclass identity recognition experiments using the LFW are more challenging due to the limited quantities of examples for many identities. This makes face recognition in the traditional multiclass sense a difficult task. To address this issue one might naturally turn to video as a plentiful source of additional imagery. The ability of video to dramatically increase the number of *in the wild* example images was one of the initial motivations for the initial version [18] of our work here. Wolf et al. have similarly created a Youtube Faces dataset for unconstrained facial images extracted from video, also noting that the source of facial image data is itself a source of variation [19].

To more accurately recognize faces in video Kim et al. used Hidden Markov Models with pose as a hidden variable [20]. In our work here we have developed a keypoint based face registration pipeline that is able to produce performance comparable to the widely used commercial alignments of the LFWa distribution. As our method is fast and there are advantages to using the same registration pipeline for all data processing, we also used it for registering faces in video here. However, we see more detailed pose modeling for our video registrations such as that of [20] as an extremely promising direction for increasing the performance of our system.

Another element of our work here is that we are particularly interested in the setting where we have a small number of labeled static face images (i.e. LFW images or other seeds), and we then wish to automatically annotate videos for the corresponding identities. As such, we focus here on the issue of capturing the uncertainly about the identities associated with faces extracted from videos automatically.

Because of the paucity of labels, the stated focus of the LFW is on the pair matching task or one-example learning, which roughly shares the same objective [17]. Pair matching using the LFW dataset is quite mature, with current accuracy of better than 88% using either a learned cosine similarity metric [21] or a model selection approach [22], both of which use an SVM to classify the pair based on a similarity metric.

Pair matching, however, is not an identical problem to the face verification task as coined by Huang et al. [17]. In face verification, an algorithm's task is to label a test image as belonging to one of a set of subjects. Although a pair matching algorithm can be used in a nearest-neighbor fashion, the standard approaches of multiclass classification can also be brought to bear. The work by Wolf et al. [23,24] is particularly representative of this work. For example, Ref. [24] specifically addresses the question of how well descriptor-based methods often used in verification tasks in object recognition work for pair matching. Wolf et al. note that for classes with a relatively large number of training examples (greater than 10) resulting in a subset of classes, quite good results can be achieved [24]. It therefore appears that the main issue is the number of labeled positives in the LFW dataset. Naturally, this raises the question of whether semi-supervised approaches can be used instead of labeling many images by hand.

Weakly labeled learning: Margin-based or margin-like properties have been presented before in the context of semi-supervised learning *i.e.* low density separation. Entropy regularization [25] finds a classifier in which the classes of the unlabeled data are maximally separated according to the entropy of the conditional distribution. Assuming the absence of a weak label, and not using a null category, the resulting objective function is similar in spirit. Transductive support vector machines (TSVMs) [26,27] find a decision boundary consistent with the labeled data which maximally separates both the labeled and test (unlabeled) data. Meanwhile, Wu et al. present the weighted margin SVM, which incorporates prior knowledge using confidence values obtained from unlabeled data [28]. However, this method also relies on a very strong low-density separation assumption as in the TSVM. TSVM is typically computationally expensive, although faster methods have been proposed [29].

The null category noise model (NCNM) is presented in [30] in the context of Gaussian processes with the goal of producing decision boundaries in regions of low density in order to produce a margin-like effect in a Bayesian framework. In the NCNM of [30] an additional target label, representing the null category, is used to enforce a low-density separation. This is achieved by restricting the model such that no data is allowed to take on the label of target label. The unlabeled data acts to “push” the decision boundary away. This model has also been extended for multiclass