# VIDEO FACE RECOGNITION WITH GRAPH-BASED SEMI-SUPERVISED LEARNING

*Effrosyni Kokiopoulou and Pascal Frossard*

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{effrosyni.kokiopoulou,pascal.frossard}@epfl.ch

**ABSTRACT**

We consider the problem of classification of multiple observations of the same object, possibly under different transformations. We view this problem as a special case of semi-supervised learning where all unlabelled examples belong to the same unknown class. We propose a low complexity solution that is able to exploit the properties of the data manifold with a graph-based algorithm. It results into a discrete optimization problem, which can be solved by an efficient algorithm. We demonstrate its performance in video-based face recognition applications, where it outperforms state-of-the-art solutions that fall short of exploiting the manifold structure of the face image data sets.

***Index Terms***— Semi-supervised learning, label propagation, video-based face recognition.

## 1. INTRODUCTION

In this work, we focus on the pattern classification problem with multiple observations that could typically represent successive frames of a video sequence. We assume that observations are produced from the same object under different transformations. In particular, the problem is to assign multiple observations of the test object $s$ to a single class of objects. We assume that we have $m$ transformed observations of $s$ of the following form

$$x_i = U(\eta_i)s,\ i = 1, \ldots, m,$$

where $U(\eta)$ denotes a (geometric) transformation with parameters $\eta$, which is applied on $s$. For instance, in the case of visual objects, $U(\eta)$ may correspond to a rotation, scaling, translation, or perspective projection of the object. We assume that each observation $x_i$ is obtained by applying a transformation $\eta_i$ on $s$, which is different from its peers (i.e., $\eta_i \neq \eta_j$, for $i \neq j$). The problem is to classify $s$ in one of the $c$ classes under consideration, using the multiple observations $x_i,\ i = 1, \ldots, m$.

Assume further that the data set is organized in two parts $X = \{X^{(l)}, X^{(u)}\}$, where $X^{(l)} = \{x_1, x_2, \ldots, x_l\} \subset \mathbb{R}^d$ and $X^{(u)} = \{x_{l+1}, \ldots, x_n\} \subset \mathbb{R}^d$, where $n = l + m$. Let also $\mathcal{L} = \{1, \ldots, c\}$ denote the label set. The $l$ examples in $X^{(l)}$ are labelled $\{y_1, y_2, \ldots, y_l\},\ y_i \in \mathcal{L}$, and the $m$ examples in $X^{(u)}$ are unlabelled. The classification problem can be formally defined as follows.

**Problem 1** *Given a set of labelled data $X^{(l)}$, and a set of unlabelled data $X^{(u)} \triangleq \{x_j = U(\eta_j)s,\ j = 1, \ldots, m\}$ that correspond to multiple transformed observations of $s$, the problem is to predict the correct class $c^*$ of the original pattern $s$.*

This problem is a particular case of semi-supervised learning [1], which generally consists in predicting the labels of $X^{(u)}$, based on the knowledge of the data points (both $X^{(l)}$ and $X^{(u)}$) and the labels of the labelled points. Note that in the generic scenario of semi-supervised learning, the test examples may belong to different classes. The above problem however presents an important additional constraint, where all the observations belong to the same class. Thus, one may view Problem 1 as a special case of semi-supervised learning, where the unlabelled data $X^{(u)}$ represent the multiple observations and they have the extra constraint that all unlabelled data examples belong to the same (unknown) class. The problem then resides in estimating the unknown class.

## 2. GRAPH-BASED CLASSIFICATION

We propose a novel graph-based algorithm built on label propagation. Label propagation methods typically assume that the data lie on a low dimensional manifold living in a high dimensional space. They rely upon the *smoothness assumption*, which states that if two data samples $x_1$ and $x_2$ are close, then their labels $y_1$ and $y_2$ should be close as well. The main idea of these methods is to build a graph that captures the geometry of this manifold as well as the proximity of the data samples. The labels of the test examples are derived by "propagating" the labels of the labelled data along the manifold, while making use of the smoothness property.

Denote by $\mathcal{M}$ the set of matrices with nonnegative entries, of size $n \times c$. Notice that any matrix $M \in \mathcal{M}$ provides a labelling of the data set by applying the following rule: $y_i = \max_{j=1,\ldots,c} M_{ij}$. We denote the initial label matrix as $Y \in \mathcal{M}$ where $Y_{ij} = 1$ if $x_i$ belongs to class $j$ and 0 otherwise. The label propagation algorithm first forms the $k$ nearest neighbor ($k$-NN) graph defined as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

where the vertices $\mathcal{V}$ correspond to the data samples $X$. An edge $e_{ij} \in \mathcal{E}$ is drawn if and only if $x_j$ is among the $k$ nearest neighbors of $x_i$.

It is common practice to assign weights on the edge set of $\mathcal{G}$. One typical choice is the Gaussian weights

$$W_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) & \text{when } (i,j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The similarity matrix $S \in R^{n \times n}$ is further defined as

$$S = D^{-1/2} W D^{-1/2}, \quad (2)$$

Effrosyni Kokiopoulou is now with the Seminar for Applied Mathematics, ETHZ, Zurich, Switzerland

This work has been partly supported by the Swiss National Science Foundation, under grant NCCR IM2.

where $D$ is a diagonal matrix with entries $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

We exploit the specificities of our particular classification problem and constrain the unknown labels to correspond to one single class. Therefore, we propose in the sequel a novel graph-based algorithm, which (i) uses the smoothness criterion on the manifold in order to predict the unknown class labels and (ii) at the same time, it is able to exploit the specificities of Problem 1.

We represent the data labels with a 1-of-$c$ encoding, which allows to form a binary label matrix of size $n \times c$, whose $i$th row encodes the class label of the $i$th example. The class label is basically encoded in the position of the nonzero element.

Suppose now that the correct class for the unlabelled data is the $p$th one. In this case, we denote by $Z_p \in R^{n \times c}$ the corresponding label matrix and we call it the $p$th *class-conditional label matrix*. Note that there are $c$ such label matrices; one for each class hypothesis. Each matrix $Z_p$ has the following form

$$Z_p = \left[ \begin{array}{c} Y_l \in R^{l \times c} \\ \hline \mathbf{1} e_p^{\top} \in R^{m \times c} \end{array} \right] \in R^{n \times c}, \quad (3)$$

where $e_p \in R^c$ is the $p$th canonical basis vector and $\mathbf{1} \in R^m$ is the vector of ones. $Z_p$ holds the labels of all data samples, assuming that all unlabelled examples belong to the $p$th class. Observe that the $Z_p$'s share the first part $Y_l$ and differ only in the second part. Since all unlabelled examples share the same label, the class labels have a special structure that reflects the special structure of Problem 1. We could then express the unknown label matrix $M$ as,

$$M = \sum_{p=1}^{c} \lambda_p Z_p, \;\; Z_p \in R^{n \times c}, \quad (4)$$

where $Z_p$ is given in (3), $\lambda_p \in \{0, 1\}$ and $\sum_{p=1}^{c} \lambda_p = 1$. In the above, $\lambda = [\lambda_1, \ldots, \lambda_c]$ is the vector of linear combination weights, which are discrete and sum to one. Ideally, $\lambda$ should be sparse with only one nonzero entry pointing to the correct class.

The classification problem now resides in estimating the proper value of $\lambda$. We propose the following objective function

$$\tilde{\mathcal{Q}}(\lambda) = \frac{1}{2} \Big( \sum_{i,j=1}^{n} W_{ij} \| \frac{1}{\sqrt{D_{ii}}} M_i - \frac{1}{\sqrt{D_{jj}}} M_j \|^2 \Big), \quad (5)$$

where the optimization variable now becomes the $\lambda$ vector. In the above, $M_i$ (resp. $M_j$) denotes the $i$th (resp. $j$th) row of $M$. In the case of normalized Laplacian, we have

$$\mathcal{Q}(\lambda) = \frac{1}{2} \sum_{i,j=1}^{n} S_{ij} \| M_i - M_j \|^2, \quad (6)$$

where $S$ is defined as in (2). It can be seen that the objective function directly relies on the smoothness assumption. When two examples $x_i$, $x_j$ are nearby (i.e., $W_{ij}$ or $S_{ij}$ is large), minimizing $\tilde{Q}(\lambda)$ and $Q(\lambda)$ results in class labels that are close too.

Thus, one may solve the optimization problem by enumerating all above possible solutions and pick the one $\lambda^*$ that minimizes $Q(\lambda)$. Then, the position of the nonzero entry in $\lambda^*$ yields the estimated unknown class. We call this algorithm **MA**nifold-based **S**moothing under **C**onstraints (MASC).

| Recognition rate (%) | MASC | MSM | KLD |
|---|---|---|---|
| $r = 4$ | 100 | 84.62 | 84.62 |
| $r = 6$ | 100 | 84.62 | 79.49 |
| $r = 8$ | 97.44 | 84.62 | 61.54 |
| $r = 10$ | 97.44 | 87.18 | 66.67 |
| $r = 12$ | 97.44 | 76.92 | 61.54 |

**Table 1**. Video face recognition results on the Honda/UCSD database.

## 3. EXPERIMENTAL RESULTS

We apply the proposed algorithm to the classification of sets of multiple images in video-based face recognition. We use the Honda/UCSD[1] database. For preprocessing we used first P. Viola's face detector [2] in order to automatically extract the facial region from each frame. Next, we downsampled the facial images to size $32 \times 32$ for computational ease. The proposed MASC method implements Gaussian weights (1) and sets $k = 5$ in the construction of the $k$-NN graph. We compare MASC to two well-known methods from the literature, namely the Mutual Subspace Method (MSM) [3] and the KL-divergence algorithm (KLD) by Shakhnarovich et al [4].

The Honda/UCSD database comes with a default splitting into training and test sets, which contains 20 training and 39 test video sequences. We use this default setup and we report the classification performance of all methods, under different data re-sampling rates. In particular, both training and test image sets are re-sampled now with step $r$ i.e., $X^{(i)} = X_i(:, 1 : r : n)$, $i = 1, \ldots, c$. Table 1 shows the recognition rates, when $r$ varies from 4 to 12 with step 2. Observe that KLD is mostly affected by $r$, by suffering loss in performance. This is not surprising since it is a density-based method and densities cannot be accurately estimated (in general) with a few samples. MSM seems to be more robust, yielding better results than KLD. Finally, MASC is again the best performer and it exhibits very high robustness against data re-sampling.

This shows the high potential in graph-based methods for efficient classification of images that belong to the same data manifold. The graph-based solution outperforms state-of-the-art subspace or statistical classification methods in video-based face recognition. Hence, this work establishes new connections between semi-supervised learning and video-based face recognition, where graph-based solutions are certainly very promising.

## 4. REFERENCES


[1] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised learning*. MIT Press, 2006.

[2] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[3] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

[4] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *European Conference on Computer Vision (ECCV)*, 3:851–868, 2002.


[1] http://vision.ucsd.edu/ leekc/HondaUCSDVideoDatabase/HondaUCSD.html