Positive and negative label propagation

Olga Zoidi[1], Anastasios Tefas[1], Nikos Nikolaidis[1], and Ioannis Pitas [1,2]

[1]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
[2]Department of Electrical and Electronic Engineering, University of Bristol, UK

***Abstract*—This paper extends the state of the art label propagation framework in the propagation of negative labels. More specifically, the state of the art label propagation methods propagate information of the form: "the sample i should be assigned the label k". The proposed method extends the state of the art framework by considering additional information of the form: "the sample i should not be assigned the label k". A theoretical analysis is presented in order to include negative label propagation in the problem formulation. Moreover, a method for selecting the negative labels in cases when they are not inherent from the data structure is presented. Furthermore, the incorporation of negative label information in two multi-graph label propagation methods is presented. Finally, a discussion on the proposed algorithm extension to out of sample data as well as scalability issues is presented. Experimental results in various scenarios showed that the incorporation of negative label information increases in all cases the classification accuracy of the state of the art.**

Keywords: label propagation, graph-based semi-supervised learning, face recognition, action recognition

I. Introduction

Label propagation is a commonly used method for classifying a set of partially labelled data by considering both the label information of the labelled data and the structure of both the labelled and unlabelled data. Most label propagation methods operate on similarity graphs [1]. In these methods, the graph nodes represent the visual data and the graph edge weights represent their pairwise similarities, which depend on the features that were selected for data representation. Then, label inference is performed along graph paths that connect labelled nodes to unlabelled ones.

The most widely used label propagation [2] performs label propagation with local and global consistency. It is essentially a manifold regularization method. For each label, one function is considered, that assigns each graph node with a real value. The initialization of the function is performed by assigning the value 1 to the nodes that are known to have a certain label and 0 to the remaining nodes. The optimization framework then regulates the function values so that, nodes with initial non-zero values maintain their original value and adjacent nodes with high weights are assigned similar values. The result of manifold regularization for a certain label on the graph nodes indicates the association of the nodes to this label. More specifically, the nodes with high function value have high association to the respective label. Finally, label assignment to the unlabelled samples is performed by selecting the label that corresponds to the function with the largest value for the respective sample node.

Label propagation is a special case of transductive semi-supervised learning. Transductive semi-supervised learning refers to the construction of classifiers that exploit class information from a set of training (labelled) data, along with the structure information of the test (unlabelled) data, in order to learn a local representation of the data space that spans on the available train and test data. As a result, such classifiers cannot be employed on "unknown" data that belong neither to the originally available training nor to test data. All label propagation methods, including transductive semi-supervised classification methods, consider information from a few training samples with known class information and the structure of all data in the training and test dataset. Even the imposition of additional discriminant constraints, in the form "manifold values on samples that belong to the same class should have small variance" and "manifold values on samples that belong to different classes should have large variance" in the optimization framework of such methods is based exclusively on the class information of the training samples. However, there are certain applications, in which additional information for the data can be exploited, that cannot be incorporated in the existing frameworks, as described in the following paragraph.

Let $\mathcal{S} = \{(\mathbf{x}_i, l_i),\ i = 1, \ldots, N | \mathbf{x}_i \in \mathbb{R}^M, l_i \in \mathcal{L}\}$ be a data set of N samples, each one belonging in one of the classes of $\mathcal{L}$, as shown in Figure 1a. The class (label) from a few samples (those with filled symbols) is known beforehand, while the class of the rest is unknown. We want to propagate the label information from the labelled data in $\mathcal{S}$ to the unlabelled ones. Let us assume that it is a priori known that the data set $\mathcal{S}$ was constructed from the union of two subsets of samples $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, shown in Figure 1(b), as follows: $\mathcal{S}_1 = \{(\mathbf{x}_i, l_i),\ i = 1, \ldots, N_1 | \mathbf{x}_i \in \mathbb{R}^M,\ l_i \in \mathcal{L}_1\}$, $\mathcal{S}_2 = \{(\mathbf{x}_{N_1+i}, l_{N_1+i}),\ i = 1, \ldots, N_2 | \mathbf{x}_{N_1+i} \in \mathbb{R}^M,\ l_{N_1+i} \in \mathcal{L}_2\}$, $N = N_1 + N_2$, $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2$, where $\mathcal{L}_1 = \{L_1, L_2, L_3\}$ and $\mathcal{L}_2 = \{L_2, L_3, L_4\}$. We notice that the set $\mathcal{S}_1$ does not contain data that belong to class L_4. Similarly, the set $\mathcal{S}_2$ does not contain data that belong to class L_1. Therefore, in order to

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grants 287674 (3DTVS) and 316564 (IMPART). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein. {tefas, nikolaid, pitas}@aiia.csd.auth.gr

optimally classify the unlabelled data in the set $\mathcal{S}$ through label propagation, apart from the state of the art label propagation assumptions, the employed framework should also consider the above mentioned observations. As a result, apart from the standard conditions:

1) the labels of the initial labelled data should be preserved and
2) data that are similar to each other should be assigned the same label,

the employed label propagation framework should be able to ensure satisfaction of the following additional conditions:

3) data that belong to the set $\mathcal{S}_1$ should not be assigned the label L_4 and
4) data that belong to the set $\mathcal{S}_2$ should not be assigned the label L_1.

Figure 1f illustrates that the application of state of the art label propagation in $\mathcal{S}$ does not take into advantage the last two conditions, therefore it does not lead to optimal classification results. Moreover, the application of state of the art label propagation separately in subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ does not lead to optimal classification results either. Figure 1c shows that the label propagation on $\mathcal{S}_1$ achieves perfect classification accuracy. However, this is not the case for the label propagation in $\mathcal{S}_2$, as shown in Figure 1d. More precisely, the samples in $\mathcal{S}_2$ belonging to classes L_2 and L_3 do not contain adequate information for the structure of the respective classes. Therefore, label propagation performance is poor for these classes. The classification results for separate label propagation on $\mathcal{S}_1$ and $\mathcal{S}_2$ are summarized in Figure 1e. On the other hand, as we will see in the following Sections, when label propagation is performed on $\mathcal{S}$ by exploiting conditions (3) and (4), optimal classification results are achieved, as illustrated in Figure 1g.

In this paper, a novel label propagation method is presented that tackles the general task of positive and negative label propagation. More specifically, the task of 'positive' label propagation tries to solve the problem of spreading the label information from a small set of data with known label to a much larger set of data with unknown label. The word 'positive' has been added in label propagation (though not existing in the literature) to distinguish between the classical ('positive') label propagation and the proposed 'positive and negative' label propagation. The ('positive') label propagation algorithm assigns the same label to data that are considered similar, according to some similarity measure. The task of negative label propagation solves the dual problem, i.e., instead of propagating the information that the i-th sample has the l-th label, we propagate the information that the i-th sample does not have the k-th label. This means that, in negative label propagation, the actual label information of the data is not known. Since negative propagation propagates label restrictions for the data, it can be considered as label constraint propagation. Experimental results on several data sets showed that the concurrent positive and negative label propagation framework has increased classification accuracy, with respect to the state of the art (positive) label propagation methods.

The rest of the paper is organized as follows. Section II presents an overview of existing works in the field of label propagation. The state of the art positive label propagation method is reviewed in Section III. An introduction to negative label propagation and its relationship to positive negative propagation is presented in Section IV. The overview of the proposed positive and negative label propagation framework is presented in Section V. The extension of the proposed positive and negative label propagation on multiple graphs is introduced in Section VI. Section IX contains the results of the conducted experiments. Finally, the conclusions are drawn in Section X.

II. Literature Overview

Label propagation methods on graphs typically define a classification function $\mathbf{f}$ on both labelled and unlabelled data that spreads the labels from labelled to unlabelled graph nodes. The classification function $\mathbf{f}$ should a) try to maintain the original labels on the labelled nodes as much as possible and b) apply the same label on unlabelled nodes that lie close to each other or belong to the same structure (e.g., cluster or manifold). The second assumption implies that $\mathbf{f}$ should be smooth over the entire graph. This results in a regularization framework of the form:

$$\min_{\mathbf{f}}\{\alpha\mathcal{C}(\mathbf{f}_L)+\beta\mathcal{S}(\mathbf{f})\}, \tag{1}$$

where $\mathcal{C}(\mathbf{f}_L)$ is a cost function on the labelled nodes that penalizes the divergence of the output labels from the initial labels and $\mathcal{S}(\mathbf{f})$ is a smoothness constraint on the whole graph. α and β are regularization parameters, which capture the trade-off between the two terms. Usually, the smoothness constraint is of the form:

$$\mathcal{S}(\mathbf{f})=\mathbf{f}^T\mathbf{S}\mathbf{f}, \tag{2}$$

where $\mathbf{S}$ is a smoothing matrix. In the majority of label propagation methods, the graph Laplacian $\mathbf{L}$ is employed as the smoothing matrix. These algorithms differ in the choice of the cost function and smoothness constraint, as well as in the incorporation of additional constraints.

In one of the earlier works, Zhou et al. [2] proposed a label propagation method with assures local and global consistency. The algorithm minimizes the quadratic cost function on the labelled data:

$$\mathcal{C}(\mathbf{f}_L)=(\mathbf{f}_L-\mathbf{Y}_L)^T(\mathbf{f}_L-\mathbf{Y}_L), \tag{3}$$

under the smoothness constraint:

$$\mathcal{S}(\mathbf{f})=\mathbf{f}^T\tilde{\mathbf{L}}\mathbf{f}, \tag{4}$$

where $\tilde{\mathbf{L}}$ is the normalized graph Laplacian. In [3], [4], the cost function is the one in (1) and the smoothness matrix is $\mathbf{S}=\mathbf{I}-\mathbf{W}$, where $\mathbf{W}$ is the weight matrix. In [5], two regularization methods are introduced, namely Tikhonov and interpolated regularization.

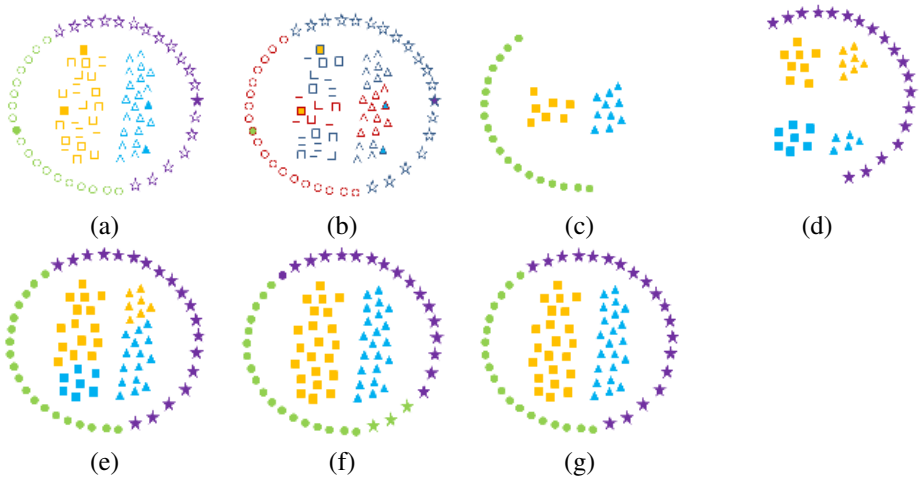


Fig. 1. a) The data in $\mathcal{S}$ that belong to four classes. The data with label L_1 are denoted with a green circle. The data with label L_2 are denoted with a yellow square. The data with label L_3 are denoted with a blue triangle. The data with label L_4 are denoted with a purple star. The data with colored filling denote the initially labeled data. b) The data in sets $\mathcal{S}_1$ and $\mathcal{S}_2$. The data in $\mathcal{S}_1$ are denoted with red border. The data in $\mathcal{S}_2$ are denoted with blue border. The data with colored filling denote the initially labeled data. c) Label propagation results on the data in $\mathcal{S}_1$. d) Label propagation results on the data in $\mathcal{S}_2$. e) Label propagation results on the data in $\mathcal{S}$ when it is performed separately on $\mathcal{S}_1$ and $\mathcal{S}_2$. f) Label propagation results on all data in $\mathcal{S}$. g) Positive and negative label propagation results on all data in $\mathcal{S}$.

The method in [6] formulates the regularization problem by defining a Gaussian Random Field on the graph and minimizing the quadratic energy function $\mathbf{f}^T\mathbf{L}\mathbf{f}$, while retaining the initial labels of the labelled nodes. The minimum energy function satisfies the harmonic property, i.e., it is equivalent to the average energy of the neighbouring nodes. Zhu et al. studied the relationship between Gaussian random fields and Gaussian processes in [7], using a spectrum transformation on the graph Laplacian matrix.

The graph mincuts method [8] targets the problem of binary label propagation with labels $\mathcal{L} = \{-1, 1\}$ as a clustering problem, which finds the minimum set of edges whose removal isolate the nodes with label 1 from those with label -1. In [9], the mincut algorithm is performed multiple times on the graph, by adding random noise on the edge weights. In each iteration, a label is assigned to the unlabelled nodes. Each unlabelled node is labelled by the label having the maximum assignment frequency. This randomized mincut algorithm provides a confidence measure for the assigned labels. In [10], spectral graph partitioning is performed through the constrained ratiocut algorithm that adds a quadratic penalty to the objective function of the standard ratio cut [11].

In cases where the data can be represented in more than one feature spaces, one graph for each representation method can be constructed. The fusion of multiple data representations can be performed either at the graph construction level (early fusion), e.g., by concatenating the separate feature vectors into a global feature vector, or at the decision level (late fusion), e.g., by learning a propagation algorithm for each data representation and fusing the propagation results. Late fusion is also called "multi-modal fusion" of "multi-modality learning" [12]. A study on early versus late fusion methods for semantic analysis of multi-modal video can be found in [13]. Label propagation methods on multiple graphs have been introduced in [12], [14], [15].

So far, we considered that the labelled and unlabelled data have a single representation. However, in many real world applications, the data can be represented in more than one feature spaces. For each representation method a new graph can be constructed. The fusion of multiple data representations can be performed either at the graph construction level (early fusion), e.g., by concatenating the separate feature vectors into a global feature vector, or on the decision level (late fusion), e.g., by learning a classification algorithm for each data representation and fusing the classification results. Late fusion is also called "multi-modal fusion" or "multi-modality learning" [12]. A study on early versus late fusion methods for semantic analysis of multi-modal video can be found in [13], where experimental results on 184 hours of video content showed that the late fusion framework had better performance for most semantic concepts, though with increased computational complexity vs the early fusion methods.

In one of the first approaches in this area, Joachims et al. [16] employed convex combinations of independent kernels. The kernels are considered independent, if they are derived from independent data representations. This method is based on the property that, any convex combination of kernels produces a new kernel. In a similar notion, a convex combination of the graph Laplacians is employed in [14], [17] and [18]. These approaches do not discriminate between graphs relevant to the classification task and more irrelevant ones, that provide no useful information. In order to alleviate this drawback, Kato et al. [15] and Wang er al. [12] proposed a propagation method that constructs a convex combination of the graph Laplacians by optimizing the weights via an iterative process, so that informative graphs are assigned larger coefficients.

First in [19] and then in [20], [21], the authors extended the single-graph regularization framework proposed in [2] in the case of multiple graphs as a weighted sum of multiple objective functions. Moreover, in [19] a sequential fusion

scheme of two graphs is proposed by sequential minimizing a two-stage optimization problem. The differences between the linear and sequential approach is in the way the similarity graphs are fussed. In the linear case, the score function $\mathbf{f}$ is spread through the information from the two graphs and, then, the results are fused. In the sequential case, first label propagation is performed based on the first similarity graph and the resulting labels are spread using the information of the second graph.

In another notion, the approach proposed in [22], regards each directed graph as a Markov chain with a unique stationary distribution similar to [23] and combines them in a mixture of Markov chains framework. In [24], 3D points and 2D images are exploited for multiple view segmentation. Three similarity graphs are constructed, which measure the 3D points similarity, the 2D color similarity and the patch histogram similarity between two joint points, i.e., vectors consisting of the coordinates of a 3D point and its corresponding patches in all image views. The final graph representing the joint similarity between two joint points is constructed by summing the three similarity graphs. In [25], multi-graph label propagation for document recommendations is performed, by fusing information of the citation matrix, the author matrix and the venue matrix. An objective function is constructed for each modality. Then, they are merged in a single objective function.

III. Positive label propagation

The task of positive label propagation tries to solve the problem of spreading the label information from a small set of data with known labels to a much larger set of data having unknown labels. Positive label propagation is simply called label propagation problem in the literature. It assigns the same label to data that are considered to be similar, according to some similarity measure. Label propagation solves the following regularization framework, introduced in [2].

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^M$ be the set of N data that belong to classes $\mathcal{L} = \{1, \ldots, L\}$. We consider that each sample belongs only to one class. We consider that the class labels $l(\mathbf{x}_i) \in \mathcal{L}$, $i = 1, \ldots, N_l$ of N_l data are known. $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ is the graph, whose edges are the data entries $\mathbf{x}_i$ in the set $\mathcal{X}$ and whose edges represent pairwise data relationships. A graph edge that connects nodes i and j is assigned with a value (similarity weight) W_{ij} that indicates the similarity between the two graph nodes. Usually, this similarity weight is computed according to the heat kernel equation [26]:

$$W(\mathbf{x}_i, \mathbf{x}_j) = W_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}\right), \tag{5}$$

where σ is the mean edge length distance among neighbors. A function $\mathbf{F} : \mathcal{X} \rightarrow \mathbb{R}^{N \times L}$ is defined, that assigns a vector of dimension L on each graph node. The vector elements represent one score value for each label. Finally, $\mathbf{Y} \in \mathbb{R}^{N \times L}$ is a matrix that represents the initial labels with entries:

$$Y_{ij} = \begin{cases} 1, & \text{if } l(\mathbf{x}_i) = l_j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Label propagation is performed by minimizing the regularization framework:

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\text{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \frac{\mu}{2}\text{tr}\left[(\mathbf{F} - \mathbf{Y})^T(\mathbf{F} - \mathbf{Y})\right], \tag{7}$$

where $\mu > 0$ is a regularization parameter and $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, $\mathbf{D} = \text{diag}_i\{\sum_{j=1:N} W_{ij}\}$ is the normalized graph Laplacian. The first term in (7) represents the clustering assumption, i.e., similar data are assigned the same label, while the second term ensures that the label of the initially labelled data remains unchanged. Minimization of $\mathcal{Q}(\mathbf{F})$ with respect to $\mathbf{F}$ leads to the following optimal solution for $\mathbf{F}^*$:

$$\mathbf{F}^* = \mu(\mathbf{L} + \mu\mathbf{I})^{-1}\mathbf{Y}. \tag{8}$$

The definition of $\mathbf{Y}$ and the clustering assumption, postulate that a high value of F^*_{ij} corresponds to a high probability that the i-th sample is assigned the j-th label. Therefore, label assignment for sample $\mathbf{x}_i$ is performed according to:

$$l_i = \arg\max_j\{F^*_{ij}\}. \tag{9}$$

IV. Negative label propagation

Negative label propagation refers to the dual problem of positive negative propagation, i.e., instead of propagating the information that the i-th sample has the l-th label, we propagate the labelling constraint that the i-th sample does not have the k-th label. This means that, in negative label propagation, the actual label information is not known. This fact renders labelling constraints as less informative than positive labels.

Label propagation is equivalent to negative propagation under the following formulation. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be the set of N data that belong to classes $\mathcal{L} = \{l_1, \ldots, l_L\}$. As in Section III, we consider that each sample belongs only to one class and that the class labels $l(\mathbf{x}_i) \in \mathcal{L}$, $i = 1, \ldots, N_l$ of N_l data are known. In terms of negative label propagation, this information is equivalent to the claim that the sample $\mathbf{x}_i$, $i = 1, \ldots, N_l$ does not have any of the labels in $\mathcal{L}^i = \mathcal{L} - \{l(\mathbf{x}_i)\}$. Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be the graph, whose nodes are the data entries $\mathbf{x}_i$ in the set $\mathcal{X}$ and whose edge weights are the pairwise data similarities W_{ij} according to the heat kernel equation (5). Let $\mathbf{\Psi} \in \mathbb{R}^{N \times L}$ be the initial state matrix, with entries:

$$\Psi_{ij} = \begin{cases} 1, & \text{if } l_j \in \mathcal{L}^i \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Negative label propagation is performed by minimizing the regularization problem defined by:

$$\mathcal{Q}(\mathbf{\Phi}) = \frac{1}{2}\text{tr}(\mathbf{\Phi}^T\mathbf{L}\mathbf{\Phi}) + \frac{\mu}{2}\text{tr}\left[(\mathbf{\Phi} - \mathbf{\Psi})^T(\mathbf{\Phi} - \mathbf{\Psi})\right], \tag{11}$$

where μ is a regularization parameter, $\mathbf{L}$ is the normalized graph Laplacian and $\mathbf{\Phi} \in \mathbb{R}^{N \times L}$ is a matrix that assigns a score on each sample for each label. Similarly to the case of positive label propagation, the optimal solution for $\mathbf{\Phi}$ is given by:

$$\mathbf{\Phi}^* = \mu(\mathbf{L} + \mu\mathbf{I})^{-1}\mathbf{\Psi}. \tag{12}$$

The definition of $\mathbf{\Psi}$ and the clustering assumption postulates that a small value of Φ^*_{ij} indicates a high probability that the i-th sample has the j-th label. Therefore, label assignment for sample $\mathbf{x}_i$ is performed according to:

$$l_i = \arg\min_j \{\Phi^*_{ij}\}. \tag{13}$$

Moreover, by definition, $\mathbf{\Psi} = \mathbf{1}_{N\times L} - \mathbf{Y}$, where $\mathbf{1}_{N\times L} \in \mathbb{R}^{N\times L}$ is a matrix of ones. By substituting $\mathbf{\Psi}$ in (13), we obtain:

$$\mathbf{\Phi}^* = \mu(\mathbf{L} + \mu\mathbf{I})^{-1}(\mathbf{1}_{N\times L} - \mathbf{Y}), \tag{14}$$

or by considering (8):

$$\mathbf{\Phi}^* = \mu(\mathbf{L} + \mu\mathbf{I})^{-1}\mathbf{1}_{N\times L} - \mathbf{F}^*. \tag{15}$$

The first term in (15) depends on the data graph and it is constant regardless the label initialization. Moreover, from (15) we notice that $\mathbf{\Phi}^*$ becomes minimum when $\mathbf{F}^*$ becomes maximum. Therefore, it can be concluded that:

$$l_i = \arg\max_j \{F^*_{ij}\} = \arg\min_j \{\Phi^*_{ij}\}. \tag{16}$$

This means that the classification results when either the positive or the negative label propagation formulation is employed are equivalent. From the above, it can be concluded that one positive label for some sample is equal to $L-1$ negative labels for the same sample, where L is the total number of labels. Only in the case of binary classification ($L = 2$) positive and negative labels have equal strength. However, even though label constraints are less informative than positive labels, their incorporation in the label propagation framework will increase its overall informativeness, as will be discussed in Subsection V.

V. Positive and negative label propagation

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^M$ be the set of N data that belong to one of the classes in $\mathcal{L} = \{l_1, \ldots, l_L\}$. We assume that two kinds of information is known for some of the samples: positive labels and negative labels (labelling constraints). Positive label information $\mathcal{P} \subset \mathcal{X} \times \mathcal{L}$ is of the form:

$$\mathcal{P} = \{(\mathbf{x}_i, l_i), i \in \{p_1, \ldots, p_q\}\}, \tag{17}$$

where the pair $(\mathbf{x}_i, l_i)$ denotes that the i-th sample has the label l_i, while negative label information $\mathcal{N} \subset \mathcal{X} \times \mathcal{L}$ is of the form:

$$\mathcal{N} = \{(\mathbf{x}_i, l'_i), i \in \{n_1, \ldots, n_m\}\}, \tag{18}$$

where the pair $(\mathbf{x}_i, l'_i)$ denotes that the i-th sample does not have the label l_i. A novel algorithm is devised that propagates both kinds of information concurrently on all samples in $\mathcal{X}$. To this end, a graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ is constructed similarly to the one in Section III. A classification function $\mathbf{F} : \mathcal{X} \to \mathbb{R}^L$ is defined on the graph nodes that assigns a real value for each label. For each label, the function should assign similar values to nodes with high similarity to each other. High F_{ij} values indicate high probability that the i-th sample has the j-th label. Finally, two matrices $\mathbf{Y}^+$ and $\mathbf{Y}^-$ are defined to represent the positive and negative label information, with entries:

$$Y^+_{ij} = \begin{cases} 1, & \text{if from prior knowledge } l(\mathbf{x}_i) = l_j \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

$$Y^-_{ij} = \begin{cases} 1, & \text{if from prior knowledge } l(\mathbf{x}_i) \neq l_j \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

By extending the regularization framework in (7), in order to incorporate the negative label information, the following objective function is defined:

$$\begin{aligned} \mathcal{Q}(\mathbf{F}) &= \frac{1}{2}\text{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + \frac{\mu}{2}\left[\mu_1\text{tr}\left((\mathbf{F}-\mathbf{Y}^+)^T(\mathbf{F}-\mathbf{Y}^+)\right)\right. \\ &- \left.\mu_2\text{tr}\left((\mathbf{F}-\mathbf{Y}^-)^T(\mathbf{F}-\mathbf{Y}^-)\right)\right]. \end{aligned} \tag{21}$$

The first term in (21) is the graph regularization term. The second term forces the initially labelled samples to retain their initial label. The third term restricts the initially negative labelled samples in obtaining the label indicated in $\mathcal{N}$. The parameter $0 < \mu < 1$ regulates the significance of the overall positive and negative label information in the optimization framework. Moreover, parameters $0 < \mu_1 < 1$ and $0 < \mu_2 < 1$ regulate the relative significance between the positive and negative label information. μ_1 and μ_2 are restricted so that $\mu_1 + \mu_2 = 1$. Since positive label information is more informative than negative label information, as discussed in Section IV, we typically choose $\mu_1 > \mu_2$. By setting the partial derivative of $\mathcal{Q}(\mathbf{F})$ with respect to $\mathbf{F}$ to zero, we obtain the following optimal solution for $\mathbf{F}$:

$$\mathbf{F}^* = \left[\mathbf{L} + (\mu\mu_1 - \mu\mu_2)\mathbf{I}\right]^{-1}\left(\mu\mu_1\mathbf{Y}^+ - \mu\mu_2\mathbf{Y}^-\right) \tag{22}$$

Finally, label assignment is performed according to (9).

Another possible straightforward approach for positive and negative label propagation could be treating the positive labels as negative ones, and then using the method introduced in Section IV. The disadvantage of this approach with respect to the proposed one is that, by combining the positive and negative label information in the same label matrix $\mathbf{\Psi}$, we assume that the significance of positive and negative label information is equivalent. However, as it will be shown in Section IXB, the significance of positive and negative labels is not equal. On the contrary, increased propagation accuracy is achieved when higher significance is given to the positive labels than to the negative ones.

Contrary to positive labels, negative label information appears more rarely in real world scenarios, e.g in person identity label propagation on facial images extracted from movies. By knowing the movie from which each facial image was extracted and the actors that appear in the cast, we can prevent a facial image from being assigned the label of an actor that does not appear in the specific movie. In the other cases, negative label information can be imposed effectively on the data manually, according to the following procedure. First, label propagation is applied on the data by considering only positive label information, according to (8) and (9). As stated before, the values in $\mathbf{F}$ are an indication on the "certainty" with which the node is assigned a label. This means that nodes,

in which the largest F_{ij} value is much larger than the second largest F_{ij} value, are more probably assigned the correct label, while nodes in which the two highest F_{ij} values are very close to each other, most probably lie in a "border" or "transition" region between two facial image classes. Label assignment to such nodes is more uncertain. The propagated labels to the nodes with the least certainty are examined, in order to form the set of negative labelled set. More specifically, for each node i, the difference between the two largest values in the i-th row of $\mathbf{F}$ is computed. The q nodes with the smallest difference value are selected and their assigned label is examined. If the label is incorrect, then the node enters the negative labelled set, describing labelling constraints. Finally, the initial state matrix $\mathbf{Y}^-$ is updated with the negative labels and label propagation is re-performed according to (22) and (9), this time considering both positive and negative label information. As it will be seen in the experiments, this choice for the negative labels increases significantly the classification accuracy of label propagation.

VI. Extension of positive and negative label propagation on multiple graphs

The proposed positive and negative label propagation framework can be easily extended to the case of label propagation on multi-modal data. In this case, multiple graphs $\mathcal{G}_1, \ldots, \mathcal{G}_K$ are constructed for the data, one for each of the K data modalities, e.g. describing color or texture or depth similarity ($K = 3$) in the case of RGB+D images. Two methods are considered for multi-graph positive and negative label propagation.

The first method that extends the Multiple Locality Preserving Projections and Cluster-based Label Propagation (MLPP-CLP) presented in [27], employs the following regularization framework:

$$\begin{aligned}\mathcal{Q}_1(\mathbf{F},\boldsymbol{\tau}) &= \frac{1}{2}\mathrm{tr}\left(\tau_k\sum_{k=1}^{K}\mathbf{F}^T\mathbf{L}_k\mathbf{F}\right)+\frac{\mu}{2}\left[\mu_1\mathrm{tr}\left((\mathbf{F}-\mathbf{Y}^+)^T\right.\right.\\ (\mathbf{F}-\mathbf{Y}^+)\big) &- \left.\left.\mu_2\mathrm{tr}\left((\mathbf{F}-\mathbf{Y}^-)^T(\mathbf{F}-\mathbf{Y}^-)\right)\right], \qquad (23)\end{aligned}$$

subject to the constraint:

$$\sum_{k=1}^{K}\tau_k = 1, \qquad (24)$$

where τ_k represents the significance of the k-th modality in information diffusion. In this method, the weights $\boldsymbol{\tau}$ are computed by the data representation method based on Multiple Locality Preserving Projections (MLPP), described in [27]. The method takes as input the multi-modal high dimensional data and calculates a single projection matrix that projects all data modalities in the same subspace of the original space. The data modalities weights for participating in the construction of the projection matrix is the same with the weight for participating in label propagation. Then, $\mathbf{F}$ is computed by setting the partial derivative of (23) with respect to $\mathbf{F}$ to zero, as follows:

$$\mathbf{F}^* = \left[\sum_{k=1}^{K}\tau_k\mathbf{L}_k + (\mu\mu_1-\mu\mu_2)\mathbf{I}\right]^{-1}\left(\mu\mu_1\mathbf{Y}^+ - \mu\mu_2\mathbf{Y}^-\right). \qquad (25)$$

The second method that extends the multi-graph label propagation algorithm (MGLP) introduced in [28] solves the following optimization problem:

$$\begin{aligned}\mathcal{Q}_2(\mathbf{F},\boldsymbol{\tau}) &= \sum_{k=1}^{K}\tau_k^2\left\{\mathrm{tr}\left[\mathbf{F}^T\mathbf{L}_k\mathbf{F}\right]\right.\\ &+ \mu\mu_1\mathrm{tr}\left[\left(\mathbf{F}-\mathbf{Y}^+\right)^T\left(\mathbf{F}-\mathbf{Y}^+\right)\right]\\ &- \left.\mu\mu_2\mathrm{tr}\left[\left(\mathbf{F}-\mathbf{Y}^-\right)^T\left(\mathbf{F}-\mathbf{Y}^-\right)\right]\right\}, \qquad (26)\end{aligned}$$

subject to the constraint (24). Sequential minimization of (26) and (24) with respect to $\mathbf{F}$ and $\boldsymbol{\tau}$ leads to the following closed form solutions:

$$\tau_k = \frac{\Lambda_k}{\sum_{k=1}^{K}\Lambda_k}, \qquad (27)$$

$$\begin{aligned}\Lambda_k &= \mathrm{tr}(\mathbf{F}^T\mathbf{L}_k\mathbf{F}) + \mu\mu_1\mathrm{tr}\left[(\mathbf{F}-\mathbf{Y}^+)^T(\mathbf{F}-\mathbf{Y}^+)\right]\\ &- \mu\mu_2\mathrm{tr}\left[(\mathbf{F}-\mathbf{Y}^-)^T(\mathbf{F}-\mathbf{Y}^-)\right]^{-1} \qquad (28)\end{aligned}$$

and

$$\mathbf{F}^* = \left[\frac{\sum_{k=1}^{K}\tau_k^2\mathbf{L}_k}{\sum_{k=1}^{K}\tau_k^2} + (\mu\mu_1-\mu\mu_2)\mathbf{I}\right]^{-1}\left(\mu\mu_1\mathbf{Y}^+ - \mu\mu_2\mathbf{Y}^-\right). \qquad (29)$$

Equations (27) and (29) are derived by setting the partial derivative of $\mathcal{Q}_2(\mathbf{F},\boldsymbol{\tau})$ with respect to $\mathbf{F}$ and $\boldsymbol{\tau}$, respectively, as in [28].

VII. Extension of positive and negative label propagation on out of sample data

The proposed positive and negative label propagation framework, described by the regularization framework in (21), can be modified in order to assign labels to out of sample data similarly to [3]. The regularization framework (21) for the sample $\mathbf{x}_i \in \mathcal{X}$ is written as:

$$\mathcal{Q}(\mathbf{f}_i) = \frac{1}{2}\sum_{j=1}^{N}W_{ij}(\mathbf{f}_i-\mathbf{f}_j)^2 + \frac{\mu}{2}\left[\mu_1(\mathbf{f}_i-\mathbf{y}_i^+)^2 - \mu_2(\mathbf{f}_i-\mathbf{y}_i^-)^2\right] \qquad (30)$$

where $\mathbf{f}_i,\ \mathbf{y}_i^+,\ \mathbf{y}_i^- \in \mathbb{R}^L$ is the i-th row of matrix $\mathbf{F}$, $\mathbf{Y}^+$ and $\mathbf{Y}^- \in \mathbb{R}^{N\times L}$, respectively. When a new sample $\mathbf{x}$ is encountered for which negative label information $\mathbf{y}^- \in \mathbb{R}^L$ is available, the smoothness criterion becomes:

$$\mathcal{Q}(\mathbf{f}(\mathbf{x})) = \frac{1}{2}\sum_{j=1}^{N}W(\mathbf{x},\mathbf{x}_j)(\mathbf{f}(\mathbf{x})-\mathbf{f}_j)^2 - \frac{\mu\mu_2}{2}(\mathbf{f}(\mathbf{x})-\mathbf{y}^-)^2. \qquad (31)$$

Since $\mathcal{Q}(\mathbf{f}(\mathbf{x}))$ is convex in $\mathbf{f}(\mathbf{x})$, it is minimized by setting $\frac{\partial\mathcal{Q}(\mathbf{f}(\mathbf{x}))}{\partial\mathbf{f}(\mathbf{x})} = 0$:

$$\mathbf{f} = \frac{1}{\sum_{j=1}^{N}W(\mathbf{x},\mathbf{x}_j) - \mu\mu_2}\left[\sum_{j=1}^{N}W(\mathbf{x},\mathbf{x}_j)\mathbf{f}_j - \mu\mu_2\mathbf{y}^-\right]. \qquad (32)$$

We notice that the optimal score vector $\mathbf{f}$ is a linear combination of the score vectors of the training data and the negative label vector. Finally, label assignment is performed according to:

$$l = \arg\max\{\mathbf{f}\}. \qquad (33)$$

VIII. Scalability

The proposed positive and negative label propagation method belongs to the Graph-based Semi-Supervised Learning (GSSL) framework. Typically, GSSL methods perform poorly on large scale data, since, only the computational complexity of the graph construction requires $\mathcal{O}(NM^2)$ computations. Several methods have been proposed for scalable GSSL methods. These methods employ approximate methods for estimating the data graph (or the graph Laplacian) and the label prediction function by considering only a subset of samples [29], [30], [31], [32], [33], [34]. Such approximate graph construction methods, as well as approximate matrix inversion approaches [35], [36] can be applied to the proposed method, in order to handle label propagation on large data. Moreover, several state of the art scalable GSSL methods based on label propagation can be straight-forwardly extended in order to incorporate negative label information, such as [37] that performs label propagation based on anchor graph regularization:

$$\min_{\mathbf{F}=[\mathbf{f}_1,\ldots,\mathbf{f}_L]} \frac{1}{2}\text{tr}\left(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F}\right) + \frac{\mu}{2}\|\mathbf{Z}_l\mathbf{F} - \mathbf{Y}_l\|_F^2, \qquad (34)$$

where $\mathbf{Z} \in \mathbb{R}^{N\times K}$ is a weight matrix that associates each sample of the N data with each one of the K anchor points and $\mathbf{Z}_l \in \mathbb{R}^{N_l\times K}$ is the submatrix that corresponds to the labelled data set $\mathcal{X}_L$. When negative label information is available, the regularization framework (34) becomes:

$$\begin{aligned}\min_{\mathbf{F}=[\mathbf{f}_1,\ldots,\mathbf{f}_L]} \frac{1}{2}\text{tr}\left(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F}\right) \quad &+ \quad \frac{\mu}{2}\left\{\mu_1\|\mathbf{Z}^+\mathbf{F} - \mathbf{Y}_l\|_F^2\right. \\ &- \quad \left.\mu_2\|\mathbf{Z}^-\mathbf{F} - \mathbf{Y}_l\|_F^2\right\}, \qquad (35)\end{aligned}$$

where $\mathbf{Z}^+$, $\mathbf{Z}^-$ are the submatrices that correspond to the positive and negative labelled data set, respectively.

IX. Experimental results

In the experiments, the performance of the proposed positive and negative label propagation framework is compared with the state of the art supervised classification methods kernel Support Vector Machines (kSVMs) [38] and k-nearest neighbours (kNNs) [39] and the state of the art label propagation methods based on local and global consistency (LP), Linear Neighbourhood Propagation (LNP) [3] and correlated label propagation (CLP) [40]. CLP is a method for propagating multiple labels that can incorporate negative label information by adding the labels $\mathcal{L}' = \{l'_1, \ldots, l'_L\}$, where label l'_i denotes that the sample is not assigned the label l_i and by considering correlations between the labels l_i and l'_j, $j = 1, \ldots, L$, $j \neq i$. Regarding the selection of the SVM kernel, we employ the heat kernel in face recognition experiments and the RBF chi-square kernel in human action recognition ones, in order to obtain the optimal classification results. Moreover, for the case of multi-graph label propagation, the proposed methods were compared to their baseline methods MLPP-CLP [27] and MGLP [28].

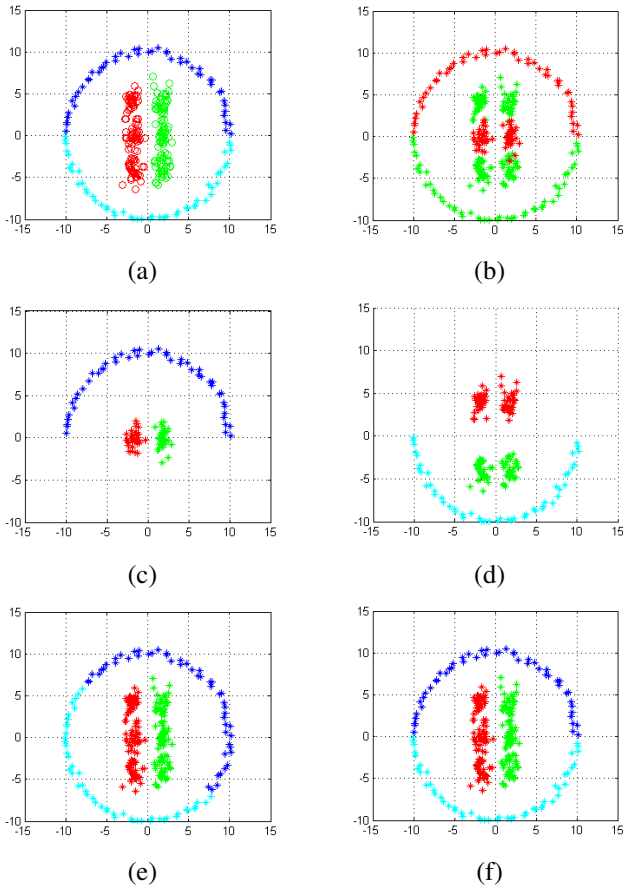


Fig. 2. Classification results for the data in experiment 1. a) The original data. b) The data sets $\mathcal{S}_1$ and $\mathcal{S}_2$. c)Positive label propagation results in $\mathcal{S}_1$. d) Positive label propagation results in $\mathcal{S}_2$. e) Positive label propagation results in $\mathcal{S}$. f) Positive and negative propagation results in $\mathcal{S}$.

A. Toy examples

1) Experiment 1: The first experiment in this section aims at the verification of the theoretic example presented in the introduction. Let $\mathcal{X}_i$, $i = 1, \ldots, 4$ the sets of samples with label L_i depicted in Figure 2a with different colours. We assume that, from prior knowledge, the data $\mathcal{S} = \bigcup_{i=1}^4 \mathcal{X}_i$ were obtained from two sets, $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, as shown in Figure 2b. Finally, we assume that the data with known labels are $\mathbf{x}_{1,20}$, $\mathbf{x}_{2,10}$, $\mathbf{x}_{31,1}$, $\mathbf{x}_{32,1}$, $\mathbf{x}_{42,1}$ and $\mathbf{x}_{43,1}$. We notice that none of the data in $\mathcal{S}_1$ has the label L_2 while none of the data in $\mathcal{S}_2$ has the label L_1. Therefore, the negative label information is of the form: "the data in $\mathbf{S}_1$ do not have the label L_1" and "the data in $\mathcal{S}_2$ do not have the label L_2". The classification accuracy of positive label propagation on the entire $\mathcal{S}$ and separately on the subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, as well as the classification accuracy of the proposed positive and negative label propagation method on $\mathcal{S}$ for $\mu_1 = 0.9$ and $\mu_2 = 0.1$ are shown in Figure 2 and Table I. We notice that the proposed positive and negative label propagation method is the only one that achieves perfect classification accuracy.

2) Experiment 2: In this experiment we test the algorithm performance for varying number of negative labels and different initialization settings. Let $\mathcal{X}_i$, $i = 1, \ldots, 4$ the sets of

TABLE I
DATA CLASSIFICATION RESULTS FOR EXPERIMENT 1.

PLP on $\mathcal{S}$	PLP on $\mathcal{S}_1$	PLP on $\mathcal{S}_2$	PLP on $\mathcal{S}_1$ and $\mathcal{S}_2$	PNLP on $\mathcal{S}$
92.95%	100%	63.70%	77.85%	100%

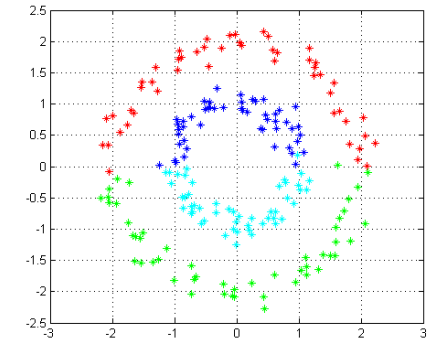
TABLE II
CLASSIFICATION RESULTS FOR THE DATA IN EXPERIMENT 3.

	classification accuracy	**computational time**
PNLP	84.90%	0.573302 sec
OSD PNLP	**85.00%**	**0.076299 sec**

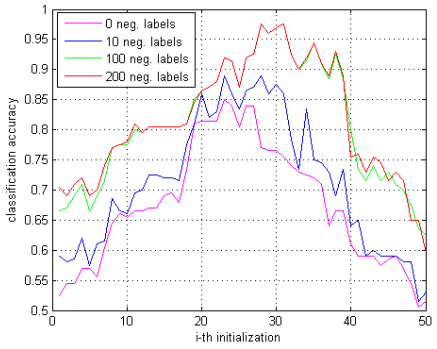
samples with label L_i that lie in two circles with radii $r_1 = 1$ and $r_2 = 1$, as shown in Figure 3a. We assume the initially labelled data set is of the form $\mathcal{S}_{L,i} = \{\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i}, \mathbf{x}_{4i}\}$, $i = 1, \ldots, 50$, i.e., we consider that positive and negative label propagation starts from 50 different initially labelled data sets. The negative labels are selected according to the method introduced in Section V. Their number varies from 10 to the total number of data, i.e., 200. The experimental results for each $\mathcal{S}_{L,i}$ and 0, 10, 100 and 200 negative labels for $\mu_1 = 0.9$ and $\mu_2 = 0.1$ are shown in Figure 3b. We notice that for each $i = 1, \ldots, 50$ the incorporation of negative label information increases the classification accuracy of label propagation. However, the increase in accuracy is not linear with respect to the number of negative labels, since the achieved classification accuracy for 100 and 200 negative labels is similar. Figure 3c depicts the average classification accuracy for various number of negative labels. It can be noticed that the increase in classification accuracy is logarithmic with respect to the number of negative labels. In the case where each and every sample is assigned one negative label the classification accuracy becomes approximately 14% larger than in the case were no negative labels are considered.

3) Experiment 3: In this experiment we test the performance of the algorithm extension to out of sample data (OSD PNLP), as described in Section VII. To this end, we employ the toy data set configuration of experiment 2 with 2,000 samples, 250 samples in each class and 200 out of sample data, 50 data in each class. The accuracy, as well as the computation time of the out of sample data classification is compared to the proposed algorithm performance if the method was re-applied on the enriched data set i.e., the original data set plus the out of sample data, and the results are shown in Table II. We notice that the algorithm extension to out of sample data achieves similar classification accuracy with the accuracy of the proposed algorithm if it was re-applied on the original plus the additional data. However, as expected, the out of sample data extension of the algorithm is 7.5 times faster than PNLP.

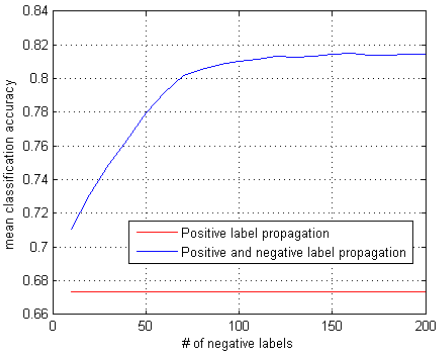
4) Experiment 4: In the final toy experiment, we test the performance of the scalable positive and negative label propagation (SPNLP) method introduced in Section VIII. To this end, we employ the toy data set configuration of experiment 2 with 2,000 samples, 250 samples in each class. The number of anchors employed in the experiment is 12, i.e., 3 anchors where uniformly selected from each class. The accuracy, as



(a)



(b)



(c)

Fig. 3. Classification results for the data in experiment 2. a) The original data. b) Positive and negative label propagation results for varying sets of initially labelled data and number of negative labels. c) Average positive and negative label propagation results for varying number of negative labels.

TABLE III
CLASSIFICATION RESULTS FOR THE DATA IN EXPERIMENT 4.

	classification accuracy	**computational time**
PNLP	**78.65%**	1.527860 sec
SPNLP	74.90%	**0.584422 sec**

well as the computation time of the PNLP and SPNLP methods are shown in Table III. We notice that the scalable PNLP achieves 3.75% lower classification accuracy than PNLP. This is due to the fact that, SPNLP is an approximate method that does not take into account the entire data information. However, as expected, SPNLP is 2.6 times faster than PNLP.

B. Face recognition

The performance of the proposed positive and negative label propagation method was tested in face recognition in two data sets: the LOST and the labelled faces in the wild (LFW) data sets.

1) LOST data set: The LOST data set [41] consists of facial images automatically extracted from 100 episodes with total duration approximately 75 hours of the tv series "LOST". The data acquisition was performed as follows. First, the Viola-Jones face detector [42] implemented in the OpenCV library was performed on each video frame, searching for facial images at various in-plane rotations and scales, obtaining approximately 100,000 facial images per episode. Since the Viola-Jones face detector returns a lot of false positives, a filtering procedure was performed on the extracted images, in order to retain only the images with high probability to actually be facial images. After this filtering procedure, approximately 10,000 facial images per episode are retained. These facial images are then organized into tracks. Finally, one facial image from each track is retained, the one with the highest probability to actually be a facial image, in order to avoid repetitive facial image instances. This results in approximately 1,000 facial images per episode.

Moreover, the LOST data set contains a set of facial images for which ambiguous label information was extracted from the screenplay. The ambiguous labels contain information about which characters appear in a certain scene. This information can be exploited in label propagation, in order to restrict the facial images that appear in the scene to be assigned only one of the character labels that are mentioned in the screenplay. It can be easily observed that this ambiguous label information consists the negative label information in the proposed positive and negative label propagation framework. Indeed, the ambiguous label information of the form: "the i-th facial image should be assigned one of the k labels $l_1, \ldots, l_{N-k}$" is equivalent to the claim: "the i-th facial image should not be assigned the labels $l_{N-k+1}, \ldots, l_N$". This data set, that was used in our experiments, consists of 1,122 facial images, belonging to 14 classes.

The experiment was performed as follows. First, the facial images, of size 90×60 pixels were cropped to 61×41 pixels and were converted to gray-scale color space. Then, the Local Binary Pattern (LBP) features with window size 5×5 were extracted for each image pixel. Finally, the Locality Preserving Projections (LPP) were applied on the facial image descriptions, in order to reduce the data dimensionality from 2,501 to 120. The performance of the proposed method was compared to that of the state of the art label propagation method [2], when 10% of the facial images were manually assigned with labels for varying values of the parameters μ_1 and μ_2. As it was pointed out in [27], the selection of the initial set of labelled images is crucial to the label propagation performance. By following the procedure introduced in [27], the initially labelled facial images were selected by clustering the facial images using a k-means algorithm. For each cluster, the facial image that lies closest to the cluster center was selected for initial labelling. Since $\mu_1 + \mu_2 = 1$, only the parameters μ and μ_2 that regulate the significance of the negative labels will be changed and μ_1 will be computed accordingly. The classification accuracy when μ and μ_2 take values in the range $[0, 1]$ is shown in Figure 5. We notice that the classification accuracy is proportional to the value of μ. Moreover, for all values of μ, the classification accuracy is the highest when μ_2 is assigned values in the range [0.05, 0.25]. For larger values of μ_2 the classification accuracy decreases significantly, especially when $\mu_2 \geq 0.45$. This is because great significance is given to negative labels that are less informative than positive labels. In the following experiments, the values of μ_1 and μ_2 are set to 0.9 and 0.1, respectively. Finally, by comparing the classification accuracy of positive label propagation achieved for $\mu_2 = 0$, which is 62.23%, with the highest classification accuracy achieved for $\mu_2 = 0.15$, which is 77.35%, we notice that the exploitation of the negative labels boosts the performance of label propagation up to approximately 15%.

Fig. 4. Examples of facial images and corresponding actor names from "LOST" series [41].

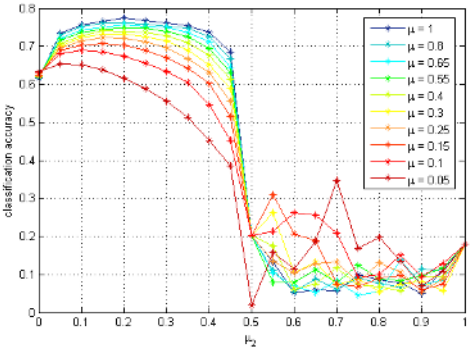


Fig. 5. Classification accuracy of positive and negative label propagation on LOST data set for varying values of μ and μ_2.

Next, we compare the performance of the proposed method to the performance of the state of the art LP, NLP, CLP, kSVM and kNNs methods. The results are shown in Table IV. We notice that the performance of the proposed method is approximately 5% better than the performance of the best state of the art method CLP that also incorporates the negative label information.

TABLE IV
CLASSIFICATION ACCURACY OF THE PROPOSED PNLP METHOD AND THE STATE OF THE ART LP, CLP, LNP, KSVM AND KNN METHODS FOR THE LOST DATA SET.

PNLP	LP	CLP	LNP	kSVM	kNN
77.35%	62.23%	72.38%	58.48%	62.77%	57.48%

Fig. 6. Sample images from the labelled Faces in the Wild data set after alignment.

2) Labelled faces in the wild data set: The labelled faces in the wild (LFW) data set [43] contains 13,233 facial images belonging to 5,749 individuals. 1,680 individuals have two or more images in the data set, while the remaining 4,069 have only one image in the data set. The images where automatically detected through the Viola-Jones face detector [42] implemented in the OpenCV library and were scaled and cropped to a fixed size of 250×250 pixels. False detections were manually erased from the data set. Finally, the facial images were aligned using the funnel algorithm [44]. Since the task of label propagation makes sense only in facial image data sets depicting individuals in multiple instances, the LFW data set was cropped, retaining only the facial images that belong to the 10 individuals with the most instances. These individuals, depicted in Figure 6 are: George W. Bush, Colin Powell, Tony Blair, Ariel Sharon, Hugo Chavez, Junichiro Koizumi, Jean Chretien, John Ashcroft, Serena Williams and Vladimir Putin. In total 1,327 facial images were retained.

The experiment was performed as follows. First, the facial images, of size 90×60 pixels were cropped to 61×41 pixels and were converted to gray-scale. Then, the Local Binary Pattern (LBP) features with window size 5×5 were extracted for each image pixel. Finally, the Locality Preserving Projections (LPP) were applied on the facial image descriptions, in order to reduce the data dimensionality from 2,501 to 75. The performance of the proposed method was compared to that of the state of the art label propagation method [2], when 10% of the facial images were manually assigned with labels through k-means clustering.

The classification accuracy of the state of the art LP and the proposed PNLP method for varying number of negative labels when the labels are selected with the method presented in Section V and when they are chosen randomly are depicted in Figure 7. In Figure 7, the classification accuracy of the state of the art LP method is the one that corresponds to zero negative labels. We notice that the incorporation of negative labels boosts the performance of label propagation up to 3.6%. More specifically, the incorporation of one negative constraint with the proposed algorithm on 1% of the data causes an

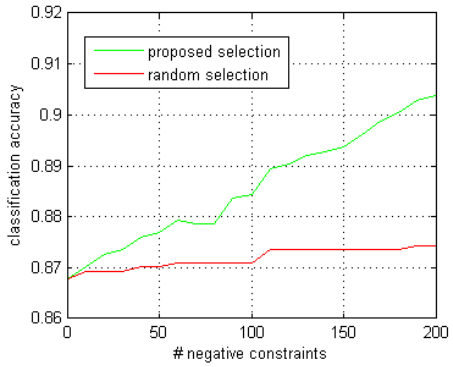


Fig. 7. Classification accuracy results of the proposed positive and negative label propagation method with the proposed and with random selection of the negative labels.

TABLE V
CLASSIFICATION ACCURACY OF THE PROPOSED PNLP METHOD AND THE STATE OF THE ART LP, CLP, LNP, KSVM AND KNN METHODS FOR THE LFW DATA SET.

PNLP	LP	CLP	LNP	SVM	k-nns
90.37%	86.77%	82.66%	83.92%	76.63%	77.81%

average increase in classification accuracy by 0.23%, while an average increase of 1% in classification accuracy is achieved by adding one negative label in 4.2% of the data. Moreover, we notice that the proposed algorithm for selecting the negative labels is much more efficient than random selection.

Next, we compare the performance of the proposed method to the performance of the state of the art LP, NLP, CLP, heat kernel SVM and kNNs methods. The results are shown in Table V. We notice that the performance of the proposed method is approximately 4.4% better than the performance of the best state of the art method LP.

C. Face recognition in Stereo Images

1) Data set description: The performance of the proposed multi-graph positive and negative label propagation methods that were presented in Section VI was tested on stereo movie data sets that consist of stereo facial images automatically extracted from three full-length stereo movies. The three stereo movies have in total 528,348 full high definition video frames of size 1080×1920 pixels and duration 6 hours, 4 minutes and 16 seconds. The data set was created as follows. First, the video shots of the stereo movies were extracted through a shot boundary detection algorithm [45]. Then, each shot was processed with an automatic face detector and an automatic face tracker algorithm, in order to extract the facial images that appear therein. The employed face detector was a modified version of the Viola-Jones frontal face detector [42] that incorporates color information [46] for eliminating a large amount of false face detections. The face detector was employed separately on each channel of the stereo video shots, retaining only the facial images that were detected in both the left and the right channel. If a facial image was detected only

in the left or the right channel, it was considered to be a false face detection and, thus, was discarded. When a facial image was detected in both channels of the video shot, it was tracked in the following 20 video frames or until the shot boundary was reached. Face tracking was performed separately on each channel through a single-channel appearance-based object tracking algorithm [47]. The tracker results in a so-called facial image trajectory consisting of facial image Regions of Interest (ROIs). This procedure was repeated for the video frames in the remaining shots. Sequential facial image trajectories that belonged to the same person and shot were concatenated into a single trajectory. In total, 171,649 facial images were detected in the three movies, forming 4,845 facial image trajectories and belonging to 129 different actors, plus some false detections. More details about the dataset can be found in [27]. Since the total number of facial images in the three movies is very large, they were reduced by sub-sampling, as follows. If the facial image trajectory contained less than 20 facial images then only the first facial image of the trajectory was selected. If the facial image trajectory contained more than 20 facial images, then one in ten facial image was selected for annotation (i.e., the 1^{st}, 10^{th}, 20^{th}, etc.). This way, more images are selected from longer trajectories. In total, 13,850 images were selected from the three movies, which represent 5.85% of the extracted facial images. The facial images were considered to belong to 131 classes, one class for each actor that appears in any of the three movies and three more that represent the false detections in each movie.

2) Experimental results: The performance of the proposed multi-graph positive and negative label propagation method (M-PNLP) based on MLPP-CLP is evaluated as follows. First, the dimensionality of the facial images is reduced by calculating a single projection matrix that preserves locality information in the left and right channel, according to the MLPP method. The facial image dimensionality is reduced from 1271 to 75. The weights $\boldsymbol{\tau}$ of each representation are calculated through the dimensionality reduction procedure. By following the procedure introduced in [27], the initially labelled facial images were selected by clustering the facial images using the k-means algorithm. For each cluster, the facial image that lies closest to the cluster center was selected for initial labelling. Then, the method described in Section V was followed, in order to select the negative labels. Finally, the classification function $\mathbf{F}^*$ was computed, according to (25). Experimental results when 5% of the facial images were initially assigned with labels for varying number of negative labels for the three movies are depicted in the green plots of Figure 8. The state of the art MLPP-CLP method is obtained for zero negative labels. Experimental results show an increase in the classification accuracy in all three movies for an increasing number of negative labels up to 2.66%.

Next, the performance of the proposed multi-graph positive and negative label propagation method (M-PNLP) based on MGLP is evaluated as follows. First, the dimensionality of the facial images is reduced by calculating a projection matrix for each data modality (i.e., the left and right channel) according to the LPP method. The facial image dimensionality in the left and right channel was reduced from 1271 to 75. By following the procedure introduced in [27], the initially labelled facial images were selected by clustering the facial images using the k-means algorithm. For each cluster, the facial image that lies closest to the cluster center were selected for initial labelling. Then, the method described in Section V was followed, in order to select the negative labels. Finally, the weights $\boldsymbol{\tau}$ and the classification function $\mathbf{F}^*$ were computed sequentially, according to (27) and (29). Experimental results, when 5% of the facial images were initially assigned with labels for varying number of negative labels for the three movies, are depicted in the red plots of Figure 8. The state of the art MGLP method is obtained for zero negative labels. Experimental results show an increase in the classification accuracy in all three movies for an increasing number of negative labels up to 2.27%. Moreover, by comparing the two proposed methods, we notice that the M-PNLP based on MLPP-CLP method achieves higher classification accuracy in all three movies.

TABLE VI
CLASSIFICATION ACCURACY OF THE PROPOSED MLPP-PNLP AND MGLP-PNLP METHODS AND THE STATE OF THE ART MLPP-CLP, MGLP, CLP, KSVM AND KNN METHODS FOR THE THREE STEREO MOVIES.

	movie 1	**movie 2**	**movie 3**
MLPP-PNLP	$\mathbf{80.65\%}$	$\mathbf{67.45\%}$	$\mathbf{68.97\%}$
GLP-PNLP	79.07%	66.93%	68.36%
MLPP-CLP	78.00%	64.79%	66.43%
MGLP	76.80%	65.41%	66.34%
CLP	59.34%	58.74%	55.46%
k-SVM	61.35%	59.19%	58.61%
kNNs	72.56%	55.88%	60.71%

Finally, we compare the performance of the proposed multi-graph methods to the performance of the state of the art MLPP-CLP, MGLP, CLP, heat kernel SVM and kNNs methods. The results are shown in Table VI. We notice that the performance of the proposed MLPP-PNLP and GLP-PNLP methods achieve the highest classification accuracy in all three data sets by approximately $2 - 2.5\%$ with respect to the best state of the art classification method.

D. Human action recognition

1) Data sets descriptions: The proposed multi-graph positive and negative label propagation methods that were presented in Section VI have been tested in the UCF11, Olympic Sports and UCF50 data sets for activity recognition. The UCF11 data set [48] consists of 1,600 Youtube videos depicting 11 action classes: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. The Olympic Sports data set [49] consists of sports videos downloaded from YouTube, depicting humans performing the following 16 sport activities: high-jump, long-jump, triple-jump, pole-vault, discus, hammer, javelin, shot put, basketball lay-up, bowling, tennis-serve, platform, springboard, snatch, clean-jerk and vault. The UCF50 data set [50] is an extension of the UCF11 data set. It

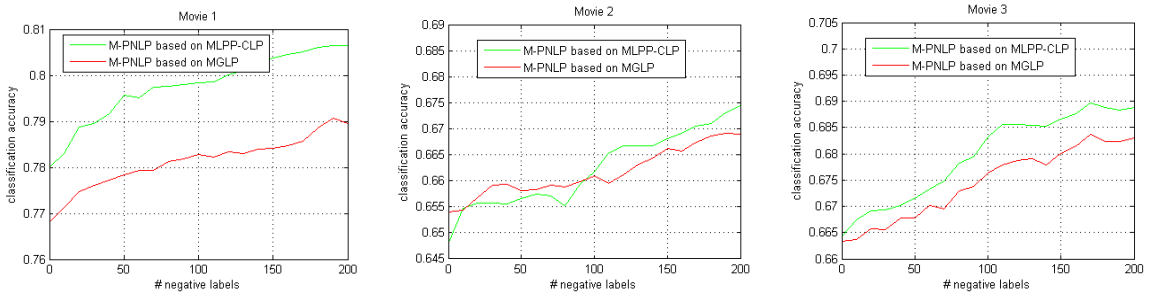


Fig. 8. Classification results of the proposed multi-graph positive and negative label propagation methods in three stereo movies.

consists of 6,680 videos downloaded from YouTube showing 50 actions: Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Lunges, Military Parade, Mixing Batter, Nun chucks, Playing Piano, Pizza Tossing, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skate Boarding, Skiing, Skijet, Soccer Juggling, Swing, Playing Tabla, TaiChi, Tennis Swing, Trampoline Jumping, Playing Violin, Volleyball Spiking, Walking with a dog, and Yo Yo. All databases are very challenging, since they consist of videos captured in completely unconstrained environments and have variations in camera motion, view point, illumination, cluttered background etc. In the UCF11 and Olympic Sports databases, each video is represented with a state of the art multi-modal action description exploiting the BoF-based video representation [51] using 5 descriptor types: Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), Motion Boundary Histograms projected on the x- and y-axis (MBHx/y) and Normalized Trajectories, evaluated on trajectories of densely sampled interest points. Each BoF representation consists of $4,000$ keywords. In the UCF50 data set, the action bank feature representation [52] was selected that consists of 205 template actions collected from all 50 action classes in UCF50 data set [50] and all six action classes from KTH database [53]. More details about the action bank feature representation can be found in [52]. It should be noted that, even though the action bank feature representation does not achieve state of the art performance, it allows us to investigate the performance of the proposed single-graph positive and negative label propagation in human action recognition.

2) Experimental results: The performance of the proposed multi-graph positive and negative label propagation methods (M-PNLP) based on MLPP-CLP and MGLP for the task of human action recognition was evaluated on the UCF11 and Olympic Sport action databased as in the previous experiment described in Section IX-C2 with the only difference that 10% of the data were initially assigned with labels. The results are depicted in Figure 9a and b. Experimental results show an increase in the classification accuracy in both data sets for an increasing number of negative labels up to 3%, for the case of MLPP-PNLP, and up to 3.59% for the case of MGLP-PNLP. Moreover, the effectiveness of the method proposed in Section V for the selection of the negative labels was also evaluated, by comparing the classification accuracy to that of random selection. The classification accuracy for random selection of the negative labels is also depicted in Figure 9a and b. We notice that the classification accuracy for the random selection of the negative labels is by far inferior to that of the proposed negative label selection method. Next, the performance of the proposed multi-graph methods was compared to the performance of the state of the art MLPP-CLP, MGLP, CLP, LNP, RBF Chi-square kernel SVM and kNNs methods. The results are shown in Table VII. We notice that the proposed methods achieve second and third best performance, after kSVM. More specifically, the average performance of kSVM is 1.4% and 1.8% better than the average performance of MLPP-PNLP and MGLP-PNLP, respectively.

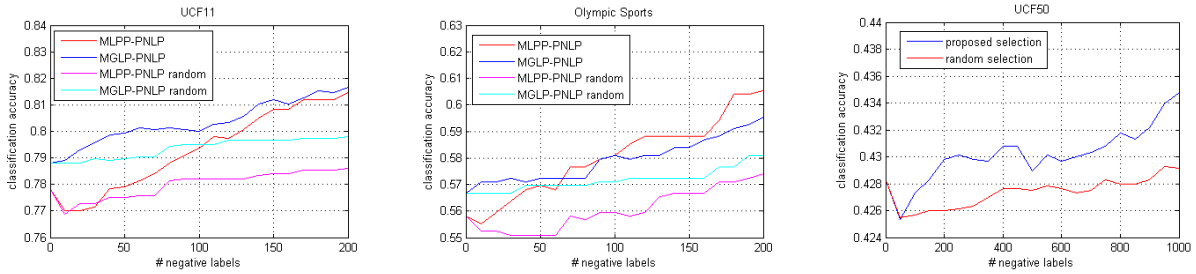
TABLE VII
CLASSIFICATION ACCURACY OF THE PROPOSED MLPP-PNLP AND MGLP-PNLP METHODS AND THE STATE OF THE ART MLPP-CLP, MGLP, CLP, LNP, KSVM AND KNN METHODS FOR THE UCF11 AND OLYMPIC SPORTS ACTION RECOGNITION DATASETS.

	ucf11	**olympic sports**
MLPP-PNLP	81.46%	60.55%
GLP-PNLP	81.67%	59.54%
MLPP-LP	77.85%	55.81%
GLP-LP	78.82%	56.67%
CLP	61.11%	42.04%
LNP	62.50%	46.05%
k-SVM	$\mathbf{83.13}\%$	$\mathbf{61.69}\%$
kNNs	62.78%	41.61%

TABLE VIII
CLASSIFICATION ACCURACY AND COMPUTATIONAL TIME (IN SECONDS) OF THE PROPOSED PNLP METHOD AND THE STATE OF THE ART LP, CLP, LNP, KSVM AND KNN METHODS FOR THE UCF50 DATABASE.

	PNLP	**LP**	**CLP**	**LNP**	**SVM**	**k-nns**
acc.	43.48%	42.83%	36.99%	40.00%	$\mathbf{45.91}\%$	40.02%
time	**33**	**33**	732	615	166	181

In the next experiment, we evaluate the performance of the proposed positive and negative label propagation method on the UCF50 action database as in the previous experiments, with the difference that in this experiment the number of selected negative labels increases from 200 to 1,000. The reason is that UCF50 consists of 50 classes, three times more classes than in the previous datasets. Therefore, the



(a) (b) (c)

Fig. 9. Classification results in human action recognition.

significance of each negative label in UCF50 is much smaller than in all previous datasets. This is verified by the experimental results in Figure 9c. We notice that the incorporation of 1,000 negative labels leads to an increase in accuracy of 0.65%. However, the accuracy is still better when the proposed negative label selection method is used over random selection. Finally, we compare the performance of the proposed method in terms on the achieved classification accuracy and the required computational time with the state of the art LP, NLP, CLP, RBF Chi-square kernel SVM and kNNs methods. The experimental results are shown in Table VIII. We notice that again the proposed method achieves the second best classification accuracy behind RBF Chi-square kernel SVM. More specifically, the performance of kSVM is approximately 2.5% better than the performance of the proposed method. However, the proposed method is 5 times faster than kernel SVM, as shown in the second row of Table VIII. It should be noted here that the measured computational time includes the time required to construct the data similarity and kernel matrices plus the time required to perform the classification.

By comparing the experimental results in all experiments in Tables IV-VIII, we notice that the proposed positive and negative label propagation framework achieves by far better performance than the state of the art methods when it is applied to the face recognition task and the second best performance when it is applied to the action recognition task. Yet, the performance of the proposed framework is close to that of the best state of the art method. On the contrary, the performance of kernel SVM, that achieves the highest classification accuracy in the action recognition task is on average 10% lower than the performance of the proposed method in the face recognition task. Moreover, when the proposed framework is applied on data with inherent negative label propagation information, such as the LOST dataset then the classification performance of the proposed framework exceeds the performance of state of the art classification methods that do not take account this information by approximately 15%. Finally, experimental results showed that the proposed positive and negative label propagation framework, along with the state of the art label propagation framework, are by far faster than the compared classification methods regarding the time required to construct the data similarity matrices and to perform the classification.

X. Conclusions

A novel method has been presented that introduces the problem of negative label propagation in the task of single-graph and multi-graph label propagation. More specifically, the state of the art label propagation methods propagate information of the form: "the sample i should be assigned the label k". The proposed method extends the state of the art framework by considering additional information of the form: "the sample i should not be assigned the label k". A theoretical analysis has been presented, in order to present the state of the art label propagation framework in the formulation of negative label propagation. Moreover, a method for selecting the negative labels in cases when they are not inherent from the data structure has been introduced. Extended experimental results in various scenarios showed that the incorporation of negative label information increases in all cases the classification accuracy of the state of the art. Moreover, the proposed positive and negative label propagation framework achieves the best and second best classification accuracy compared to state of the art supervised and label propagation methods when applied to the tasks of face recognition and human action recognition, respectively. The effectiveness of the proposed framework becomes more significant when the data contain inherent negative label information.

References

[1] X. Zhu, *Semi-Supervised Learning Literature Survey*. Technical Report, University of Wisconsin - Madison, 2008.

[2] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004, pp. 321–328.

[3] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. ACM, 2006, pp. 985–992.

[4] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by knn-sparse graph-based label propagation over noisily tagged web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 14:1–14:15, 2011.

[5] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *COLT*. Springer, 2004, pp. 624–638.

[6] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, pp. 912–919.

[7] X. Zhu, J. Lafferty, and Z. Ghahramani, "Semi-supervised learning: From gaussian fields to gaussian processes," School of CS, CMU, Tech. Rep., 2003.

[8] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 19–26.

[9] A. Blum, J. Lafferty, M. R. Rwebangira, and R. Reddy, "Semi-supervised learning using randomized mincuts," in *Proceedings of the twenty-first international conference on Machine learning*, ser. ICML '04. ACM, 2004, pp. 13–.

[10] T. Joachims, "Transductive learning via spectral graph partitioning," in *Proceedings of the international conference on Machine learning*, ser. ICML '03. ACM, 2003, pp. 290–297.

[11] L. Hagen and A. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 11, no. 9, pp. 1074 –1085, sep 1992.

[12] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733–746, 2009.

[13] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05, 2005, pp. 399–402.

[14] A. Argyriou, M. Herbster, and M. Pontil, "Combining graph laplacians for semi-supervised learning," in *Advances in Neural Information Processing Systems 18*. MIT Press, 2005, pp. 67–74.

[15] T. Kato, H. Kashima, and M. Sugiyama, "Robust label propagation on multiple networks," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 35 –44, 2009.

[16] T. Joachims, T. J. De, N. Cristianini, and N. R. A. Uk, "Composite kernels for hypertext categorisation," in *Proceedings of the International Conference on Machine Learning (ICML*. Morgan Kaufmann Publishers, 2001, pp. 250–257.

[17] K. Tsuda, H. Shin, and B. Schölkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. 2, pp. 59–65, 2005.

[18] V. Sindhwani and P. Niyogi, "A co-regularized approach to semi-supervised learning with multiple views," in *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

[19] H. Tong, J. He, M. Li, C. Zhang, and W. Ma, "Graph based multi-modality learning," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 862–871.

[20] M. Wang, X. Hua, X. Yuan, Y. Song, and L. Dai, "Optimizing multi-graph learning: towards a unified video annotation scheme," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 862–871.

[21] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733–746, 2009.

[22] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07. ACM, 2007, pp. 1159–1166.

[23] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. ACM, 2005, pp. 1036–1043.

[24] J. Xiao, J. Wang, P. Tan, and L. Quan, "Joint affinity propagation for multiple view segmentation," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–7.

[25] D. Zhou, S. Zhu, K. Yu, X. Song, B. Tseng, H. Zha, and C. Giles, "Learning multiple graphs for document recommendations," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 141–150.

[26] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[27] O. Zoidi, A. Tefas, N. Nikolaidis, and I. Pitas, "Person identity label propagation in stereo videos," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1358–1368, Aug 2014.

[28] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733–746, 2009.

[29] O. Delalleau, Y. Bengio, and N. Le Roux, "Efficient non-parametric function induction in semi-supervised learning," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 96–103.

[30] X. Zhu and J. Lafferty, "Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 1052–1059.

[31] I. W. Tsang and J. T. Kwok, "Large-scale sparsified manifold regularization," in *Advances in Neural Information Processing Systems*, 2006, pp. 1401–1408.

[32] K. Zhang, J. T. Kwok, and B. Parvin, "Prototype vector machine for large scale semi-supervised learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1233–1240.

[33] R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised learning in gigantic image collections," in *Advances in neural information processing systems*, 2009, pp. 522–530.

[34] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 679–686.

[35] M. Wu, B. Yin, A. Vosoughi, C. Studer, J. R. Cavallaro, and C. Dick, "Approximate matrix inversion for high-throughput data detection in the large-scale mimo uplink," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2155–2158.

[36] F. Soleymani, "On a fast iterative method for approximate inverse of matrices," *Communications of the Korean Mathematical Society*, vol. 28, no. 2, pp. 407–418, 2013.

[37] W. Liu, J. Wang, and S.-F. Chang, "Robust and scalable graph-based semisupervised learning," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2624–2638, 2012.

[38] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[39] Y. Wu, K. Ianakiev, and V. Govindaraju, "Improved k-nearest neighbor classification," *Pattern recognition*, vol. 35, no. 10, pp. 2311–2318, 2002.

[40] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1719–1726.

[41] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 919–926.

[42] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[43] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[44] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *ICCV*, 2007.

[45] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[46] G. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for automatic face detection and tracking," in *Visual Communications and Image Processing 2005*. International Society for Optics and Photonics, 2005, pp. 59 602C–59 602C.

[47] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491–1506, 2004.

[48] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1996–2003.

[49] J. Niebles, C. Chend, and F.-F. L., "Modeling temporal structure of decomposable mition segemnts for activity classification," in *European Conference on Computer Vision*, 2010.

[50] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[51] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[52] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1234–1241.

[53] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.