# Person identity label propagation in stereo videos

Olga Zoidi, Anastasios Tefas, Nikos Nikolaidis, Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 54124, GREECE
tel: +30 2310 996361
{tefas, nikolaid, pitas}@aiia.csd.auth.gr

***Abstract*—In this paper a novel method is introduced for propagating person identity labels on facial images extracted from stereo videos. It operates on image data with multiple representations and calculates a projection matrix that preserves locality information and a priori pairwise information, in the form of must-link and cannot-link constraints between the various data representations. The final data representation is a linear combination of the projections of all data representations. Moreover, the proposed method takes into account information obtained through data clustering. This information is exploited during the data propagation step in two ways: to regulate the similarity strength between the projected data and to indicate which samples should be selected for label propagation initialization. The performance of the proposed Multiple Locality Preserving Projections with Cluster-based Label Propagation (MLPP-CLP) method was evaluated on facial images extracted from stereo movies. Experimental results showed that the proposed method outperforms state of the art methods.**

## I. Introduction

Annotation typically aims at multimedia data archival and fast search, based on their semantic annotation (tags). This situation arises, for example, in the case of television content annotation in broadcasters' audiovisual archives. In this case, archivists usually perform a coarse annotation of the entire video, which, in many cases, is insufficient for journalists to directly access video shots/frames of interest. Such problems can be overcome with semi-automatic annotation techniques, based on label propagation [1], which is a semi-automatic process for spreading semantic labels from a small set of available labeled data to a much larger set of unlabeled data. In the case of television content annotation, one type of semantic information, which is of interest to archivists, concerns the person identities and appearances in videos to be archived. Label propagation techniques take into consideration the following assumptions: 1) visual data, e.g., video shots, frames, facial images, that are similar to each other, according to a similarity measure, or that lie in the same feature space structure (e.g., cluster, manifold) should be assigned the same label and 2) the initial labeled data should retain their label during/after label propagation.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). 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.

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 labeled nodes to unlabeled ones. When the data can be represented in multiple ways, more than one graphs can be constructed to represent the same data set. Such a situation arises for example in the case of label propagation in stereo videos, where the visual information is coming from the two luminance channels, or from the video and the depth (disparity) channels [2]. In such cases, the label propagation methods can take into consideration information extracted from all similarity graphs.

Label propagation effectiveness depends on two factors: the graph construction and the label inference method. Graph construction deals with the choice of the data representation and the pairwise similarity (or distance) metric. In the case where the data are facial images, i.e., detected rectangular facial image regions of interest (ROI) of size $N_x \times N_y$ pixels, the image data usually undergo a dimensionality reduction step. A typical reduced dimensionality image representation can be obtained by finding a projection matrix that maps the images $\mathbf{x}_i$, $i = 1, \ldots, M$ residing on the original space $\Re^N$, $(N = N_x N_y)$ on a subspace $\Re^L$, $L << N$. The $N \times L$ projection matrix $\mathbf{A}$ forms the basis matrix in the resulting space $\Re^L$ and the data projections $\mathbf{x}'_i = \mathbf{A}^T \mathbf{x}_i \in \Re^L$ form the image features to be used in graph construction. Such a subspace representation that is widely used in person identification algorithms is the Locality Preserving Projection (LPP) [3]–[5]. In LPP, the data are projected to such a reduced dimensionality space, so that the locality information of the original data is preserved, i.e., when $\|\mathbf{x}_i - \mathbf{x}_j\|$ is small, then $\|\mathbf{x}'_i - \mathbf{x}'_j\|$ is small as well. In [6], sparsity constraints were imposed in the objective function of LPP, so that the sparse reconstructive weights are preserved, while in [5], a regularized LPP method is presented, that extracts useful discriminant information from the entire feature space. LPP is an unsupervised dimensionality reduction technique, since it exploits information obtained only from the data structure in the original space $\Re^N$. However, several extensions of LPP have been proposed, that exploit prior information about the data and extend LPP to the semi-supervised and the fully-supervised framework. Discriminant constraints obtained from

the availability of labeled data are introduced in [7], [8]. In [7], the maximization of the between-class distance and the minimization of the within-class distance is proposed. The discriminant LPP method proposed in [8] maximizes the difference between the locality preserving between-class scatter matrix and locality preserving within-class scatter matrix. Moreover, orthogonality constraints were imposed on the discriminant LPP in [9].

After data feature extraction through LPP and graph construction through a chosen similarity measure, label propagation is performed on the visual data (graph nodes), according to a label inference method, which specifies the way the labels are spread from the set of labeled data to the set of unlabeled data. Usually, iterative label inference methods are employed [1]. In these algorithms, label spread is performed gradually on the unlabeled data, according to some update rule. The final label allocation converges to a stationary state, as $t \rightarrow \infty$. The stationary state of the iterative algorithm can be computed beforehand. Therefore, in such cases, these methods are performed in a single step. Such label propagation methods are introduced in [10]–[12]. The performance of label propagation methods depends highly on the selection of the initially labeled data set. A method for selecting the initially labeled data set is presented in [2].

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" [13]. A study on early versus late fusion methods for semantic analysis of multi-modal video can be found in [14]. Label propagation methods on multiple graphs have been introduced in [13], [15], [16].

In this paper, we propose a novel method for label propagation on data with multiple representations that finds application in person identity label propagation on multi-view camera systems. The proposed method exploits information obtained from multiple data representations, by finding a projection matrix that preserves locality information and additional a priori pairwise information between the data in all data representations. The method calculates one projection matrix for all data representations. Each data representation influences the projection matrix calculation with a weight that is learned automatically from the regularization framework. The data representations are then projected on the same reduced-dimensionality space. The projections of the various data representations are then combined in two ways, in order to perform label propagation. In the first way, the data are combined in a single representation as a weighted linear combination of the various data representations. Label propagation is then performed like any label propagation method that operates on a single graph. In the second way, one graph is constructed for each data representation projection. Then, label propagation is performed concurrently on the graphs by extending the single-graph regularization framework to consider information obtained from all graphs. Moreover, the proposed method exploits the data structure in label inference in two novel ways: in the construction of the weight matrix of the final data representation and in the selection of the data set which will be initially manually labeled. The proposed method was employed on facial images extracted from stereo movies. In this case, the facial images have inherently two representations, one for the left and one for the right stereo video channel. The scope for the algorithm is to propagate facial identity information on all the facial images that appear in the videos, starting from a small set of manually labeled facial images. Experimental results showed the effectiveness of the proposed method in propagating face identity information with respect to state of the art methods.

The rest of the paper is organized as follows. An overview on locality preserving projections and label propagation with local and global consistency is presented in Section II. Locality preserving projections on multiple graphs are introduced in Section III. Two methods for fusing the various data representation projections, as well as a method for initializing the labeled data set is introduced in Section IV. Experimental results are presented in Section V. Finally, conclusions are drawn in Section VI.

## II. Locality Preserving Projections and Label Propagation

### A. Locality Preserving Projections

Locality Preserving Projections [3] is a method for linear dimensionality reduction that operates on graphs. LPP calculates a projection matrix that projects the data to a reduced dimensionality space, so that the locality information of the original data is preserved. Let $\mathcal{X} = \{\mathbf{x}_i \in \Re^N, i = 1, \ldots, M\}$ be the data set in the original space and $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be the graph, whose nodes are the data entries $\mathbf{x}_i$ in the set $\mathcal{X}$ and whose edges are the pairwise data relationships. The edge in the graph that connects the nodes $i$ and $j$ is assigned with a value $W_{ij}$ that indicates the similarity between the adjacent graph nodes. This similarity is often computed according to the heat kernel equation [3]:

$$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}}, \tag{1}$$

were $\sigma$ is the mean edge length distance among neighbors. LPP tries to find a projection matrix $\mathbf{A} \in \Re^{N \times L}$, such that, if the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ is strong (i.e., $W_{ij}$ takes a large value), then the data projections $\mathbf{x}'_i = \mathbf{A}^T\mathbf{x}_i$, $\mathbf{x}'_j = \mathbf{A}^T\mathbf{x}_j$, are mapped close to each other. Let $\mathbf{a} \in \Re^N$ be a transformation vector (a column of $\mathbf{A}$). The objective of LPP is to find the vectors $\mathbf{a}$ that minimize the function:

$$\sum_{i,j=1}^{M} (\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j)^2 W_{ij} = \mathbf{a}^T\mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T\mathbf{a} = \mathbf{a}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{a}, \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \Re^{N \times M}$ is the data matrix, $\mathbf{D}$ is a diagonal matrix with entries $D_{ii} = \sum_{j=1}^{M} W_{ij}$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian, subject to the constraint:

$$\mathbf{a}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{a} = 1, \tag{3}$$

which guarantees distance preservation in $\Re^L$. The transformation vector that minimizes (2) is the eigenvector that corresponds to the minimum eigenvalue of the generalized eigenvector problem:

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{a} = \lambda \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{a}. \tag{4}$$

Finally, the projection matrix $\mathbf{A}$ consists of the eigenvectors that correspond to the $L$ smallest eigenvalues of (4).

### B. Locality Preserving Projections with Pairwise Constraints

LPP can also incorporate side information, in the form of pairwise similarity and dissimilarity constraints, in order to improve classification performance [17]. In this case, besides locality preservation, the objective of LPP is to find a projection matrix such that the data projections satisfy the pairwise constraints, i.e., data that satisfy similarity constraints should be mapped close to each other and data that satisfy dissimilarity constraints should be mapped far away from each other. Let $\mathcal{S}$ be the set of similar pairs:

$$\mathcal{S} = \{(i,j) | \mathbf{x}_i, \mathbf{x}_j \text{ must have the same label}\}, \tag{5}$$

and $\mathcal{D}$ be the set of dissimilar pairs:

$$\mathcal{D} = \{(i,j) | \mathbf{x}_i, \mathbf{x}_j \text{ must have different labels}\}. \tag{6}$$

Two weight matrices are constructed, $\mathbf{W}_s$ and $\mathbf{W}_d$, for the similar and dissimilar constraints, respectively, as follows:

$$W_{s,ij} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{S} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$$W_{d,ij} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{D} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The objective function of LPP then becomes:

$$\arg\min_{\mathbf{a}} \sum_{i,j=1}^{M} (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j)^2 W_{ij} + \sum_{(i,j) \in \mathcal{S}} (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j)^2 - \sum_{(i,j) \in \mathcal{D}} (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j)^2, \tag{9}$$

or equivalently:

$$\arg\min_{\mathbf{a}} \mathbf{a}^T \mathbf{X} (\mathbf{D} + \mathbf{D}_s + \mathbf{D}_d - \mathbf{W} - \mathbf{W}_s - \mathbf{W}_d) \mathbf{X}^T \mathbf{a}, \tag{10}$$

where $\mathbf{D}_s$, $\mathbf{D}_d$ are diagonal matrices with entries $D_{s,ii} = \sum_{j=1}^{M} W_{s,ij}$ and $D_{d,ii} = \sum_{j=1}^{M} W_{d,ij}$, subject to the constraint:

$$\mathbf{a}^T \mathbf{X} (\mathbf{D} + \mathbf{D}_s + \mathbf{D}_d) \mathbf{X}^T \mathbf{a} = 1. \tag{11}$$

### C. Propagating the Pairwise Constraints

Intuitively, we can say that, if we know that two nodes have the same labels from prior knowledge, then the neighbors of these nodes should also have the same label, due to neighboring node similarity. In a similar argumentation, if we know that two nodes have dissimilar labels, then the nodes that belong to the neighborhood of one node should have different label from the other node and vice versa. This means that we can generalize the pairwise constraints to include neighboring nodes in an iterative procedure, similarly to label propagation. Let $\mathcal{N}_i$ be the neighborhood of node $i$, based on, e.g., thresholding the Euclidean distance between two nodes and $\mathbf{P} \in \Re^{M \times M}$ be a sparse weight matrix with entries:

$$P_{ij} = \begin{cases} \frac{1}{|\mathcal{N}_i|}, & \text{if } j \in \mathcal{N}_i \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where $|\mathcal{N}_i|$ is the cardinality of the set $\mathcal{N}_i$. It is clear that the sum of each row of $\mathbf{P}$ is 1. We define a function $\mathbf{F}_s$ that assigns a real value to every graph node that indicates its label similarity to the other graph nodes. In each iteration, the node incorporates some information from its neighbors and retains some information from its initial state $\mathbf{W}_s$. At $t$-th iteration, the label similarity is equal to:

$$\mathbf{F}_s^{(t)} = a \mathbf{P} \mathbf{F}_s^{(t-1)} + (1-a) \mathbf{W}_s, \tag{13}$$

or equivalently:

$$\mathbf{F}_s^{(t)} = (a\mathbf{P})^{(t-1)} \mathbf{W}_s + (1-a) \sum_{t'=1}^{(t-1)} (a\mathbf{P})^{(t')} \mathbf{W}_s. \tag{14}$$

Parameter $a$, $0 \leq a \leq 1$, regulates the percentage of information the node will receive from its neighbors and from its initial state. Since $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$, the Perron-Frobenius theorem [18] holds and (14) converges to the steady state [19]:

$$\mathbf{F}_s = (1-a)(\mathbf{I} - a\mathbf{P})^{-1} \mathbf{W}_s. \tag{15}$$

Similarly, the label dissimilarity is propagated according to:

$$\mathbf{F}_d^{(t)} = a \mathbf{P} \mathbf{F}_d^{(t-1)} + (1-a) \mathbf{W}_d, \tag{16}$$

which converges to the steady state:

$$\mathbf{F}_d = (1-a)(\mathbf{I} - a\mathbf{P})^{-1} \mathbf{W}_d. \tag{17}$$

### D. Label Propagation with Local and Global Consistency

Let us define the set of labeled data $\mathcal{X}_L = \{\mathbf{x_i}\}_{i=1}^{m_l}$, which are assigned labels from the set $\mathcal{L} = \{l_j\}_{j=1}^{Q}$ and a set of unlabeled data $\mathcal{X}_U = \{\mathbf{x_i}\}_{i=1}^{m_u}$. Without loss of generality, we define the set of labeled and unlabeled data as $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_l}, \mathbf{x}_{m_l+1}, \ldots, \mathbf{x}_M\}$, $M = m_l + m_u$. The vector $\mathbf{Y} = [y_1, \ldots, y_{m_l}, 0, \ldots, 0]^T = [\mathbf{Y}_L^T | \mathbf{Y}_U^T]^T \in \mathcal{L}^M$ contains the labels of the labeled data in the first $m_l$ positions and takes the value 0 in the last $m_u$ positions. The objective of label propagation methods is to spread the labels in $\mathcal{L}$ from the set of labeled data $\mathcal{X}_L$ to the set of unlabeled data $\mathcal{X}_U$.

A label propagation algorithm that exploits local and global consistency was introduced in [12]. It begins with the construction of a graph, whose nodes are the labeled and unlabeled data and whose edges are the pairwise data relationships, based on, e.g., the heat kernel equation (1). Then, a set of functions $f_i$, $i = 1, \ldots, M$ is defined that assigns on the $i$-th graph node one value for every possible label. By defining the matrix $\mathbf{F} = [\mathbf{f}_1^T, \ldots, \mathbf{f}_M^T]^T \in \Re^{M \times Q}$, label propagation is performed by the iterative process [12]:

$$\mathbf{F}^{t+1} = a\mathbf{S}\mathbf{F}^t + (1-a)\mathbf{Y}, \tag{18}$$

where:

$$\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} \tag{19}$$

and $\mathbf{D}$ is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$. Essentially, the matrix $\mathbf{S}$ represents how much the label value of a node is affected from the label of its neighboring nodes. The matrix $\mathbf{Y} \in \Re^{M \times Q}$ represents the initial state, with values:

$$Y_{ij} = \begin{cases} 1, & \text{if node } i \text{ is labeled as } y_i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

It is proven in [12] that the iterative procedure (18) converges to the solution:

$$\mathbf{F} = (1-a)(\mathbf{I} - a\mathbf{S})^{-1}\mathbf{Y}. \tag{21}$$

Finally, the label information is propagated to the nodes according to the following decision rule:

$$y_i = \arg\max_j F_{ij}. \tag{22}$$

It is proven in [12] that the iterative process given by (18) is equivalent to a manifold regularization problem defined by:

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\text{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + \mu\text{tr}\left((\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y})\right), \tag{23}$$

where $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-1/2}$ is the normalized graph Laplacian and $\mu = \frac{1-a}{a}$.

## III. Locality Preserving Projections on Multiple Graphs

In this paper, we propose a novel method for performing linear dimensionality reduction on data with multiple representations, by satisfying additional pairwise similarity and dissimilarity constraints, called Multiple-graph Locality Preserving Projections (MLPP). The proposed method searches for a $N \times L$ projection matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_L]$ that operates on all visual data views (e.g., the left/right video channel) and also searches for the optimal linear combination of the data projections. Let $\mathbf{x}_{ki}$, $k = 1, \ldots, K$ be the different data representations of facial image $i$ and $\mathbf{a}_l$, $l = 1, \ldots, L$ the projection vectors that form the columns of the projection matrix. The objective of the proposed method is the minimization of the function:

$$\begin{aligned} \arg\min_{\mathbf{a}_l, \boldsymbol{\tau}} \sum_{k,l} \tau_k \Bigg\{ & \sum_{i,j=1}^{M} (\mathbf{a}_l^T\mathbf{x}_{k,i} - \mathbf{a}_l^T\mathbf{x}_{k,j})^2 W_{k,ij} \\ & + \beta \sum_{(i,j)\in\mathcal{S}} (\mathbf{a}_l^T\mathbf{x}_{k,i} - \mathbf{a}_l^T\mathbf{x}_{k,j})^2 F_{s,ij} \\ & - \gamma \sum_{(i,j)\in\mathcal{D}} (\mathbf{a}_l^T\mathbf{x}_{k,i} - \mathbf{a}_l^T\mathbf{x}_{k,j})^2 F_{d,ij} \Bigg\} + \varepsilon\|\boldsymbol{\tau}\|^2, \end{aligned} \tag{24}$$

subject to the constraints:

$$\mathbf{a}_l^T\mathbf{a}_j = \delta_{lj}, \;\; \sum_k \tau_k = 1, \;\; \tau_k \geq 0, \;\; l,j = 1, \ldots, L, \;\; k = 1, \ldots, K, \tag{25}$$

where $\tau_k$, $k = 1, \ldots, K$ is the weight of the $k$-th data representation in the optimization framework, $\beta$, $\gamma$ are parameters that regulate the significance of the pairwise similarity and dissimilarity constraints, respectively and $\varepsilon$ is a regularization parameter that prevents the coefficients vector $\boldsymbol{\tau}$ from taking increased value for only one image representation. The first sum in (24) ensures that the locality information of the data in the original space is preserved in the projected space. The second/third sum in (24) ensure that the similar/dissimilar data pairs are projected close to/away from each other. Finally, the first constraint in (25) ensures that the projection matrix $\mathbf{A}$ is orthonormal. By simple algebraic manipulations, (24) can be written as:

$$\arg\min_{\mathbf{a}_l, \boldsymbol{\tau}} \sum_{k,l} \tau_k \mathbf{a}_l^T \mathbf{X}_k \left(\mathbf{L}_k + \beta\mathbf{L}_s - \gamma\mathbf{L}_d\right) \mathbf{X}_k^T \mathbf{a}_l + \varepsilon\|\boldsymbol{\tau}\|^2, \tag{26}$$

where $\mathbf{L}_k = \mathbf{D}_k - \mathbf{W}_k$ is the graph Laplacian for the $k$-th data representation and $\mathbf{L}_s = \mathbf{D}_s - \mathbf{F}_s$, $\mathbf{L}_d = \mathbf{D}_d - \mathbf{F}_d$ are the graph Laplacians of the pairwise similarity and dissimilarity constraints, respectively. $\mathbf{L}_k$ varies according to the data representation, while $\mathbf{L}_s$, $\mathbf{L}_d$ are constant for all representations.

By selecting the parameters $\beta$, $\gamma$ so that the matrix $\mathbf{L}_c = \beta\mathbf{L}_s - \gamma\mathbf{L}_d$ is positive semi-definite, the cost function (26) under the constraints (25) is convex, with respect to the variables $\mathbf{a}_l$ and $\boldsymbol{\tau}$. Indeed, if the data number $M$ is greater than the data dimensions $N$, then we can consider that the data are linearly independent and the rank of $\mathbf{X}_k \in \Re^{N \times M}$ is $N$. From the properties of semi-definite matrices, if $\mathbf{L}_c \in \Re^{M \times M}$ is positive semi-definite and $\mathbf{X}_k \in \Re^{N \times M}$ is of rank $N$, then $\mathbf{X}_k\mathbf{L}_c\mathbf{X}_k^T$ is positive semi-definite [20]. Moreover, the sum of positive semi-definite matrices is also a positive semi-definite matrix. The Hessian of the quadratic function (26) with respect to $\mathbf{a}_l$ is a positive semi-definite matrix. Moreover, the Hessian of the quadratic function (26) with respect to $\boldsymbol{\tau}$ is the identity matrix, which is positive definite. Therefore, the quadratic problem defined in (26) is convex with respect to either $\mathbf{a}_l$ or $\boldsymbol{\tau}$. In the experiments, we set $\beta = 100$ and $\gamma = 0.01$ to ensure that matrix $\mathbf{L}_c$ is positive semi-definite. The optimization problem is solved iteratively for $\mathbf{a}_l$ and $\boldsymbol{\tau}$ as follows:

1) First, $\boldsymbol{\tau}$ is initialized with the values $\tau_k = \frac{1}{K}$, $k = 1, \ldots, K$.

2) The system (25), (26) is solved for $\mathbf{a}$ by constructing the Lagrangian function:

$$\mathcal{L}(\mathbf{a}_l, \lambda) = \mathbf{a}_l^T \left[ \sum_k \tau_k \mathbf{X}_k \left( \mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d \right) \mathbf{X}_k^T \right] \mathbf{a}_l - \lambda \mathbf{a}_l^T \mathbf{a}_l. \quad (27)$$

By setting the partial derivative of the Lagrangian function with respect to $\mathbf{a}_l$ equal to zero $\frac{\partial \mathcal{L}(\mathbf{a}_l, \lambda)}{\partial \mathbf{a}_l} = 0$, we get:

$$\left[ \sum_k \tau_k \mathbf{X}_k \left( \mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d \right) \mathbf{X}_k^T \right] \mathbf{a}_l = \lambda \mathbf{a}_l. \quad (28)$$

It is easy to see that the projection vectors $\mathbf{a}_l$, $l = 1, \ldots, L$ that minimize the objective function are the eigenvectors that correspond to the $L$ smallest eigenvalues of matrix $\sum_k \tau_k \mathbf{X}_k \left( \mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d \right) \mathbf{X}_k^T$. Finally, the projection matrix $\mathbf{A}$ is constructed: $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_L]$.

3) Next, (25), (26) are solved with respect to $\boldsymbol{\tau}$, for the projection matrix $\mathbf{A}$ that was calculated as in (28). By writing (26) in matrix form with respect to $\boldsymbol{\tau}$, we get:

$$\arg\min_{\tau} \sum_k \tau_k \text{tr} \left[ \mathbf{A}^T \mathbf{X}_k (\mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d) \mathbf{X}_k^T \mathbf{A} \right] + \varepsilon \boldsymbol{\tau}^T \boldsymbol{\tau}, \quad (29)$$

subject to the constraints:

$$\boldsymbol{\tau}^T \mathbf{1}_K = 1, \;\; \tau_k \geq 0, \;\; k = 1, \ldots, K, \quad (30)$$

where $\mathbf{1}_K \in \Re^K$ is a vector of ones. The system (29)-(30) is a quadratic programming problem with respect to $\boldsymbol{\tau}$ and can be solved with any quadratic programming solver.

4) Steps 2 and 3 are repeated until convergence.

The convergence of the proposed iterative procedure is proved as follows. Let $G(\mathbf{A}^{(t)}, \boldsymbol{\tau}^{(t)})$ be the value of the objective function (24) at iteration $t$. Since the parameters $\beta$, $\gamma$ are chosen in such a way, so that the matrix $\mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d$, $k = 1, \ldots, K$ is positive semi-definite, $G(\mathbf{A}^{(t)}, \boldsymbol{\tau}^{(t)})$ is convex with respect to either $\mathbf{A}$ or $\boldsymbol{\tau}$. This means that, the solution of (28) $\mathbf{A}^{(t+1)}$ produces the minimum value of the objective function $G(\mathbf{A}^{(t)}, \boldsymbol{\tau}^{(t)})$ given the value of $\boldsymbol{\tau}^{(t)}$ and, subsequently, the solution of (29)-(30) $\boldsymbol{\tau}^{(t+1)}$ produces the minimum value of the objective function $G(\mathbf{A}^{(t+1)}, \boldsymbol{\tau}^{(t)})$ given the value of $\mathbf{A}^{(t+1)}$, or equivalently [21]:

$$G(\mathbf{A}^{(t)}, \boldsymbol{\tau}^{(t)}) \geq G(\mathbf{A}^{(t+1)}, \boldsymbol{\tau}^{(t)}) \geq G(\mathbf{A}^{(t+1)}, \boldsymbol{\tau}^{(t+1)}). \quad (31)$$

Since $G(\mathbf{A}, \boldsymbol{\tau}) \geq 0$ and it is convex with respect to either $\mathbf{A}$ or $\boldsymbol{\tau}$, (31) implies that $G(\mathbf{A}, \boldsymbol{\tau})$ converges asymptotically to a local minimum.

After the projection matrix $\mathbf{A}$ and the coefficients vector $\boldsymbol{\tau}$ are computed, the data projections $\mathbf{X}'_k$ of representation $k$ to the reduced dimensional space are computed as:

$$\mathbf{X}'_k = \mathbf{A}^T \mathbf{X}_k. \quad (32)$$

The data projections $\mathbf{X}'_k$ are then fused in order to perform label propagation, as will be described in Section IV. The steps for dimensionality reduction through locality preserving projections on multiple graphs is summarized in Figure 1.

## IV. Label Propagation on Multiple Graphs

After the computation of the data projections $\mathbf{X}'_k$, $k = 1, \ldots, K$, label propagation is performed on the projected data by fusing the information obtained from all representations. The fusion of multiple data representations can be performed either at the graph construction level (early fusion) or at the decision level based on label propagation (late fusion).

### A. Early Fusion

In early fusion, the data representation projections are linearly combined into a single data representation, by:

$$\mathbf{X}' = \sum_k \tau_k \mathbf{X}'_k, \quad (33)$$

where the weights $\boldsymbol{\tau}$ are the ones computed by MLPP (Section III). The new data representations (33) are used to construct a new data graph, having a weight matrix computed from (1). Label propagation is then performed like any label propagation method that operates on a single graph, by exploiting local and global consistency, according to (18).

### B. Late Fusion

In late fusion, one graph is constructed for each data representation projection $\mathbf{X}'_k$, $k = 1, \ldots, K$ with weight matrices $\mathbf{W}_k$ computed from (1). Then, label propagation is performed concurrently on the $K$ graphs, by extending the single-graph regularization framework (23) as a weighted sum of $K$ objective functions:

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2} \sum_{k=1}^{K} \tau_k \text{tr} \left( \mathbf{F}^T \mathbf{L}_k \mathbf{F} \right) + \frac{1}{2} \mu \text{tr} \left( (\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y}) \right), \quad (34)$$

where $\mathbf{L}_k$ is the normalized graph Laplacian of representation $k$. The weights $\tau_k$ are determined as in Section III. The regularization framework (34) is similar to the regularization framework proposed in [22] with the difference that, in the proposed method, the weights $\tau_k$ have been pre-computed during the data dimensionality reduction. (34) is convex with respect to $\mathbf{F}$. Therefore, the global optimum can be found by setting the partial derivative of $\mathcal{Q}(\mathbf{F})$ equal to zero:

$$\frac{\partial \mathcal{Q}(\mathbf{F})}{\partial \mathbf{F}} = \sum_k \tau_k \left( \mathbf{L}_k \mathbf{F} \right) + \mu (\mathbf{F} - \mathbf{Y}) = 0. \quad (35)$$

The global optimum is then given by:

$$\mathbf{F} = (1 - a) \left( \mathbf{I} - a \sum_k \tau_k \mathbf{S}_k \right)^{-1} \mathbf{Y}, \quad (36)$$

where we set $\mathbf{L}_k = \mathbf{I} - \mathbf{S}_k$, $\mathbf{S}_k = \mathbf{D}^{-1/2} \mathbf{W}_k \mathbf{D}^{-1/2}$ and $a = \frac{1}{1+\mu}$.

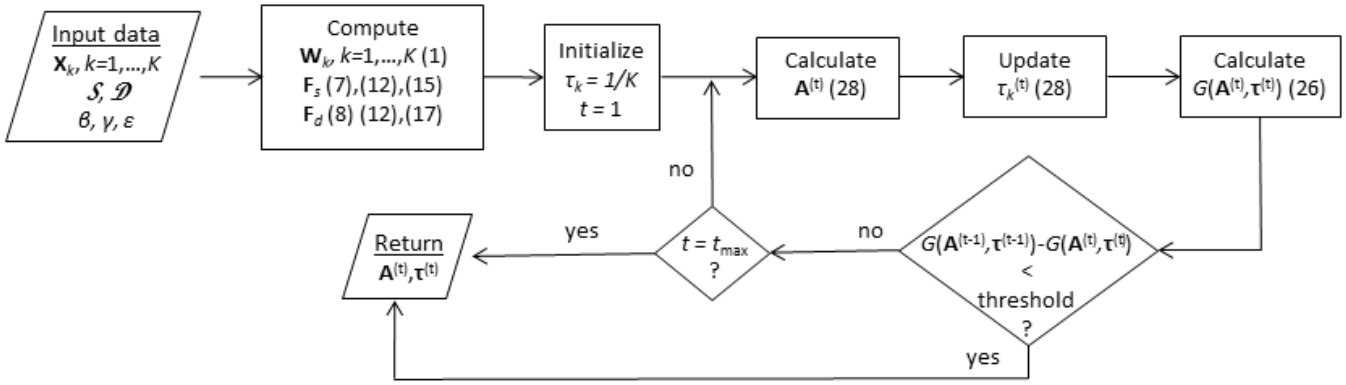


Fig. 1. MLPP algorithm flowchart (the numbers in parentheses refer to equations in the text).

### C. Iterative Label Propagation Initialization

It was observed that label propagation performance depends highly on the initialization of the propagation algorithm, i.e., on the selection of the initial data samples that will be manually assigned a label. Therefore, a more structured procedure for selecting the initial labeled data set should be followed. Such a structured procedure is introduced in this paper.

The proposed label propagation method exploits the prior information obtained from the data structure, by applying a clustering algorithm, e.g., k-means clustering, or n-cut [23]. First, the data $\mathbf{X}'$ (33) are divided into clusters and, then the data graph weight matrix $\mathbf{W}$ (or $\mathbf{W}_k$ for the case of late fusion) is calculated according to (1). In order to increase label propagation between samples in the same cluster and punish propagation between samples in different clusters, the entries of the weight matrix (1) are modified as follows:

$$W_{ij} = \begin{cases} W_{ij}, & \text{if nodes } i,j \text{ belong to the same cluster} \\ \eta W_{ij}, & \text{if nodes } i,j \text{ belong to different clusters,} \end{cases} \tag{37}$$

where $0 \leq \eta \leq 1$ is a penalizing parameter. By setting $\eta = 0$, label propagation between different clusters is prohibited. On the other hand, if we set $\eta = 1$, then we get the standard label propagation algorithm (18)-(21). The influence of the parameter $\eta$ selection and of the selected cluster number is examined in Section V.

Instead of using the same penalizing parameter for propagation between clusters that are close to each other and clusters that are further away, we introduce a second method for recalculating the weight matrix, which takes into account the distance between the cluster centers. More specifically, we define a new weight matrix $\mathbf{W}'$ with entries:

$$W'_{ij} = \begin{cases} W_{ij}, & \text{if nodes } i,j \text{ belong to the same cluster} \\ \zeta(c_i,c_j)W_{ij}, & \text{if nodes } i,j \text{ belong to different clusters,} \end{cases} \tag{38}$$

where $W_{ij}$ is given by (1), $c_i$, $c_j$ are the clusters of nodes $i$ and $j$, respectively, and:

$$\zeta(c_i,c_j) = e^{-\frac{\|\mathbf{x}_{c_i}-\mathbf{x}_{c_j}\|^2}{\sigma}}, \tag{39}$$

where $\mathbf{x}_{c_i}$, $\mathbf{x}_{c_j}$ are the centers of clusters $c_i$, $c_j$, respectively and $\sigma$ is defined as in (1). Then, from each cluster, the node with the highest within-cluster degree centrality [24] is selected to be in the set $\mathcal{X}_L$ of initially labeled samples. The within-cluster degree centrality for a node $i$ that belongs to cluster $c$ is measured by summing the edge weights that connect the node $i$ with all other nodes of cluster $c$ $d_i = \sum_{j\in\mathcal{N}_c} W_{ij}$, where $\mathcal{N}_c$ the cardinality of cluster $c$. Intuitively, the node with the highest within-cluster degree centrality is the most representative cluster node, i.e., the node with the highest similarity to all other cluster nodes, according to heat kernel similarity (1). Then, the initial state matrix $\mathbf{Y}$ is constructed and label propagation is performed according to (18)-(22).

The values in $\mathbf{F}$ (22) are an indication on the "certainty" with which the node is assigned a label, i.e., nodes in which the highest $F_{ij}$ value is much larger to the second highest $F_{ij}$ value are more probable to be 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' region between two visual data classes. Label assignment to such nodes is more uncertain. The nodes which were assigned a label with the least certainty form the next set of nodes that will be manually labeled and inserted in the set $\mathcal{X}_L$ of labeled nodes. More specifically, for each node $i$ we compute the difference between the two largest values in the $i$-th row of $\mathbf{F}$. The $q$ nodes with the smallest difference value are inserted in the set $\mathcal{X}_L$ and, the initial state matrix $\mathbf{Y}$ is updated, in order to include the newly manually labeled nodes and label propagation is performed again, according to (18)-(22) (or (36) for the case of late fusion). The procedure is repeated and the labeled set $\mathcal{X}_L$ is enriched with $q$ nodes at the time with the smallest $p_i$ value, until the cardinality of the set $\mathcal{X}_L$ is a determined percentage (e.g., $5\%$) of the overall data number. The steps for the iterative label propagation initialization method are summarized in Figure 2.

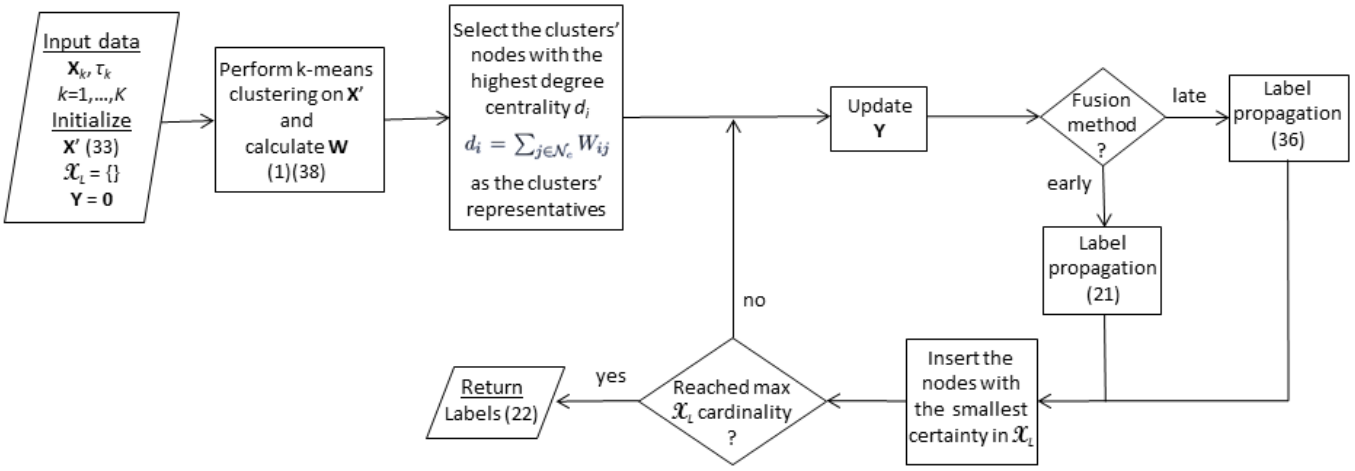


Fig. 2. Iterative label propagation initialization flowchart (the numbers in parentheses refer to equations in the text).

## V. Experiments

### A. Stereo Facial Image Database Description

Experimental results were conducted on three stereo movies. The task was to perform person identity (label) propagation on the facial images that appear in these movies with a procedure that emulates the annotation procedure followed in television archives by archivists upon new content ingestion. All these movies have full high definition video frame size ($1080 \times 1920$ pixels), total duration 6 hours, 4 minutes and 16 seconds and 528,348 frames in total.

First, the movies were processed with a shot cut detection algorithm and the shot boundaries were detected. Then, the facial images were automatically extracted by performing automatic face detection and tracking. The face detector used was the Viola-Jones face detector [25], modified to incorporate color information [26] that eliminates a large amount of false detections. Face detection was performed separately on the left and right video channels, retaining only the facial images that were detected in both channels. When a facial image was detected in both channels, it was tracked for the next 20 frames or until a shot cut was detected, using a single channel appearance-based object tracker [27]. The tracker results in a so called facial image trajectory consisting of facial image ROIs. The procedure was repeated for the remaining video frames. Sequential facial image trajectories that belonged to the same person were concatenated into a single trajectory. In total, 171,649 facial images were detected forming 4,845 facial image trajectories and belonging to 129 different actors plus some false detections. More details about the characteristics of each movie can be found in Table I. Since the number of the extracted facial images is very large, the resulting graph weight matrix of the facial images would be very large and too expensive to compute. In order to decrease the computational complexity and increase annotation speed, we make the following assumptions for the data:

1) facial images that belong to the same facial image trajectory belong to the same actor,
2) facial images appearing in the same video frame belong to different actors.

According to the first assumption, only one image from each trajectory, e.g. the first one, is required for the actor identity annotation process. The remaining images in the facial image trajectory simply adopt the label of the first image. However, by selecting only one image from each trajectory, we discard information about the trajectory length during the propagation procedure. In order to retain this information, we select more images from the longer trajectories and less from the shorter ones. More precisely, if the facial image trajectory contains less than 20 facial images then only the first facial image of the trajectory was selected. If the facial image trajectory contains more than 20 facial images, then every 10 facial images of the trajectory one facial image was selected for annotation (i.e., the $1^{st}$, $10^{th}$, $20^{th}$, etc.). 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.

The two previously mentioned assumptions on facial image ROIs form automatically extracted prior knowledge that is exploited in the label propagation procedure in the form of pairwise similarity and dissimilarity constraints, discussed in Section II-B. More specifically, the similarity and dissimilarity weight matrices (7) - (8) are constructed as:

$$W_{s,ij} = \begin{cases} 1, & \text{if images } i,j \text{ are in the same trajectory} \\ 0, & \text{otherwise,} \end{cases} \tag{40}$$

$$W_{d,ij} = \begin{cases} 1, & \text{if images } i,j \text{ are in the same frame} \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

In total 9,003 pairwise similarity and 2,034 pairwise dissimilarity constraints were automatically extracted. A more detailed description on the constructed data set can be found in Table II. Finally, the facial images are aligned with the

TABLE I
SUMMARIZATION OF MOVIE AND FACIAL IMAGE PROPERTIES

| | **Duration** | **video frame number** | **No. of facial image ROIs** | **No. of facial image trajectories** | **No. of actors** | **Facial image dataset size** |
|---|---|---|---|---|---|---|
| Movie 1 | 02 : 06 : 21 hours | 181,763 | 64,717 | 1,532 | 26 | 5,398 |
| Movie 2 | 01 : 44 : 31 hours | 150,361 | 44,625 | 1,435 | 44 | 3,498 |
| Movie 3 | 02 : 16 : 24 hours | 196,224 | 62,307 | 1,878 | 58 | 4,954 |
| Total | 06 : 04 : 16 hours | 528,348 | 171,649 | 4,845 | 128 | 13,850 |

TABLE II
FACIAL IMAGE DATA SET INFORMATION

| | **Actor classes** | **dataset size** | **Number of similarity constraints** | **Number of dissimilarity constraints** |
|---|---|---|---|---|
| movie 1 | 27 | 5,398 | 3,866 | 857 |
| movie 2 | 45 | 3,498 | 2,063 | 260 |
| movie 3 | 59 | 4,954 | 3,074 | 917 |
| total | 131 | 13,850 | 9,003 | 2,034 |

funnel algorithm [28], which was also used in the Labeled Faces in the Wild [29] data set.

### *B. Effect of parameters $\eta$ and $\zeta$ to MLPP-CLP*

In this section, the effect of the prior information imposed on the projected data weight matrix obtained through clustering in the classification performance of the proposed MLPP-CLP algorithm is examined. More specifically, the classification performance was tested for parameter $\eta$ values 0 (i.e., label propagation between clusters is prohibited), 0.2, 0.4, 0.6, 0.8 and 1 (i.e., no clustering information is exploited in the weight matrix). The number of clusters used in the experiments was 150, 50 and 100, for the Movies 1-3, respectively and the left and right channel representations were fused with the early fusion method. In all experiments, the data dimensionality is reduced to 75. The experimental results are shown in the first six rows of Table III, where it is evident that the optimal classification accuracy for each movie is achieved for $\eta = 0.6$ or $\eta = 0.8$. Table III shows that the classification accuracy of the algorithm is sensitive to the selection of $\eta$. More specifically, we notice that the classification accuracy increases for values of $\eta$ between 0.4 and 0.8, with respect to the case in which no clustering information is taken into account ($\eta = 1.0$). On the contrary, when $\eta$ takes values less than $0.4$, i.e., there is a strong prohibition for label propagation between clusters, the classification accuracy decreases with respect to the case where $\eta = 1.0$. The decrease is maximum for $\eta = 0$, i.e., when label propagation is restricted within the clusters. Finally, we examine the classification performance when the exponential parameter $\zeta$ in (39), that takes into account the distance between the cluster centers, is exploited in label propagation. We notice that, even though the use of $\zeta$ does not lead to the best classification accuracy for any of the three movies, it still increases the classification accuracy with respect to the state of the art ($\eta = 1.0$) in two out of three movies. In the rest of the experiments, clustering information will be imposed to the weight matrix with the adaptive parameter $\zeta$.

TABLE III
SIGNIFICANCE OF PARAMETERS $\eta$ AND $\zeta$ TO THE CLASSIFICATION ACCURACY OF MLPP-CLP

| | **Movie 1** | **Movie 2** | **Movie 3** | **Average** |
|---|---|---|---|---|
| $\eta$ = 0.0 | 73.69% | 55.33% | 60.08% | 64.20% |
| $\eta$ = 0.2 | 76.37% | 65.53% | 67.97% | 70.64% |
| $\eta$ = 0.4 | 77.67% | 65.66% | 68.34% | 71.31% |
| $\eta$ = 0.6 | 78.12% | **67.96%** | **68.48%** | **72.11%** |
| $\eta$ = 0.8 | **79.08%** | 66.43% | 67.44% | 71.73% |
| $\eta$ = 1.0 | 77.51% | 65.42% | 67.06% | 70.73% |
| $\zeta$ | 76.63% | 66.54% | 67.46% | 70.81% |

TABLE IV
SIGNIFICANCE OF DATA INITIALIZATION TO THE CLASSIFICATION ACCURACY OF MLPP-CLP

| | **Random initialization** | | **Proposed initialization** | |
|---|---|---|---|---|
| | Early Fusion | Late Fusion | Early Fusion | Late Fusion |
| Movie 1 | 75.69% | 75.98% | 79.26% | 80.52% |
| Movie 2 | 59.09% | 58.12% | 67.10% | 67.10% |
| Movie 3 | 57.22% | 62.00% | 68.38% | 68.39% |
| Average | 64.11% | 65.70% | 71.76% | 72.20% |

### *C. Effect of labeling initialization to MLPP-CLP*

In this section, we examine the effect of the proposed method for selecting the initially labeled data set in the classification accuracy. First, the facial images of the movies 1-3 were divided into 170, 100 and 150 clusters, respectively and the facial images that correspond to the cluster centers were manually labeled. In each iteration of the algorithm described in Subsection IV-C 33, 25 and 33 images in the border between clusters (that have the smallest "certainty") were assigned labels manually. The procedure was repeated 3 times. In all experiments, the initial manually labeled data set consists 5% of the facial images, i.e., for the movies 1-3 the manually labeled data set consists of 270, 175 and 250 facial images, respectively.

First, we compare the classification performance, when the initially labeled data set is selected randomly, without taking into account any prior information, as typically proposed in the literature. Experimental results are shown in the first and third rows of Table IV for the random and the proposed initialization method, respectively. We notice that the proposed initialization method causes a high increase in the classification accuracy in all three videos, that results in an increase of the average classification accuracy by 5.37%. Moreover, by comparing the classification accuracy of the proposed method using early and late fusion of the left and right channels data representation we notice that, the late fusion of the data representations achieves slightly better classification accuracy.

Next, we test how the classification accuracy changes in each iteration of the initialization procedure described in Subsection IV-C. The change in the classification accuracy with every iteration for the three movies is shown in Figure 3a-c. We notice that, as expected, the classification accuracy increases as more data are inserted into the initially labeled data set up to $5\%$ of the overall facial image data set. However, the increase in accuracy is not constant in each iteration, but depends on the set of data that enter the initially labeled set. For example, when early fusion is used, in Movie 1 (Figure 3a) there is a greater increase in the third algorithm iteration than in the first two, in Movie 2 (Figure 3b) the maximum increase in accuracy is observed in he second iteration and in Movie 3 (Figure 3c) an almost constant increase in accuracy is observed in all three iterations. Similar conclusions can be drawn when late fusion is used. In overall, for the first movie (Figure 3a) the classification accuracy between the first and final iteration increases by approximately $3\%$, in the second movie (Figure 3b) by $4.5\%$ and in the third movie (Figure 3c) by $3.5\%$.

### D. Comparison of MLPP to other subspace methods

The performance of the proposed Locality Preserving Projections on multiple graphs (MLPP) with early (MLPP-E) and late (MLPP-L) fusion of the stereo information, with the parameters used in Subsection V-C, is compared to the performance of similar state of the art subspace techniques, namely the standard Locality Preserving Projections (LPP) [3], Orthogonal Locality Preserving Projections (OLPP) [9], Locality Preserving Projections with Pairwise Constraints (PCLPP) [17], Graph-Optimized Locality Preserving Projections (GoLPP) [4] and Neighborhood Preserving Embedding (NPE) [30], with the standard (random) initialization. In order to test the significance of the stereo information to the classification accuracy, we compared the performance of the proposed algorithm to the performance of LPP, OLPP, PCLPP, GoLPP and NPE when they operate on one luminance channel of the stereo video. The experimental results are shown in Table V. We notice that in all three videos the classification accuracy of the proposed MLPP-E and MLPP-L algorithms achieve a much better classification accuracy. The average increase in accuracy with MLPP with respect to the best single-channel subspace method PCLPP is $8.15\%$.

Next, we test the performance of the single-channel subspace methods, when they operate separately on the left and right channels of the stereo videos and the late fusion method described in [22] is employed for performing label propagation on the stereo facial images. The experimental results are shown in Table VII. We notice that, when the existing dimensionality reduction techniques are combined with the label propagation approach and a late fusion approach they increase the classification performance with respect to single channel label propagation, yet the performance is worse than the performance of the proposed MLPP-E and MLPP-L algorithms. More specifically, the average increase in accuracy with MLPP with respect to the best state of the art stereo method is $5.7\%$.

### E. Algorithm performance on data with more modalities

The proposed method has been tested in the UCF11 data set [31], that consists of 1,600 Youtube videos depicting 11 action classes. Each video is represented with the state of the art action description exploiting the BoF-based video representation [32] evaluated on 5 descriptor types, each description type consisting one data modality ($K = 5$): 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. In the experiments, $10\%$ of the action videos were manually assigned labels. The dimension of the data is reduced from 1000 to 75. Since no prior information is available for the data set in the form of pairwise constraints, the matrices $\mathbf{L}_s$, $\mathbf{L}_d$ are set equal to $\mathbf{0}$. The experimental results for the proposed method and the state of the art method LPP, which achieved the best performance, are shown in Table VII. We notice that, the performance of MLPP-CLP with late fusion is $15.32\%$ better than the performance of the best single-modality LPP and $6.48\%$ better than the multi-channel LPP. By comparing the results for $K = 2$ and $K = 5$, we notice that, when the proposed algorithm is employed on data with more modalities, we obtain a greater increase in the classification accuracy with respect to the single-modality methods. This is because the different types of information obtained from more modalities complement one another and thus, increase the classification accuracy. Moreover, by comparing the performance of the proposed early and late fusion methods we notice that, when the data modality is small ($K = 2$), then the late fusion method is marginally better than the early fusion. The supremacy of late fusion is more evident when the data modality number increases.

### F. Parameter Selection

As mentioned in Section III, the parameters $\beta$ and $\gamma$ are chosen in such a way, so that the matrix $\mathbf{L}_c = \beta\mathbf{L}_s - \gamma\mathbf{L}_d$ is positive semi-definite. The positive semi-definiteness of $\mathbf{L}_c$ is required, so that the objective function is convex with respect to the projection matrix $\mathbf{A}$, i.e., the optimization framework converges to a global minimum. In practice, this is achieved by selecting the value of $\beta$ to be larger that the value of $\gamma$ by several orders of magnitude. In our experiments we set $\beta = 100$ and $\gamma = 0.01$, i.e., $\beta$ is larger than $\gamma$ by 4 orders of magnitude. The parameter $\varepsilon$ regulates the values of the coefficients vector $\tau$. When $\varepsilon$ takes small values, i.e., when the value of the first term (double sum) in (26) is larger than the value of the second term (the norm of the coefficient vector $\boldsymbol{\tau}$), then $\tau$ takes the value 1 for one data representation and 0 for all others, i.e., only one data representation method is taken into account during dimensionality reduction. This is undesired, since the purpose is to exploit information from

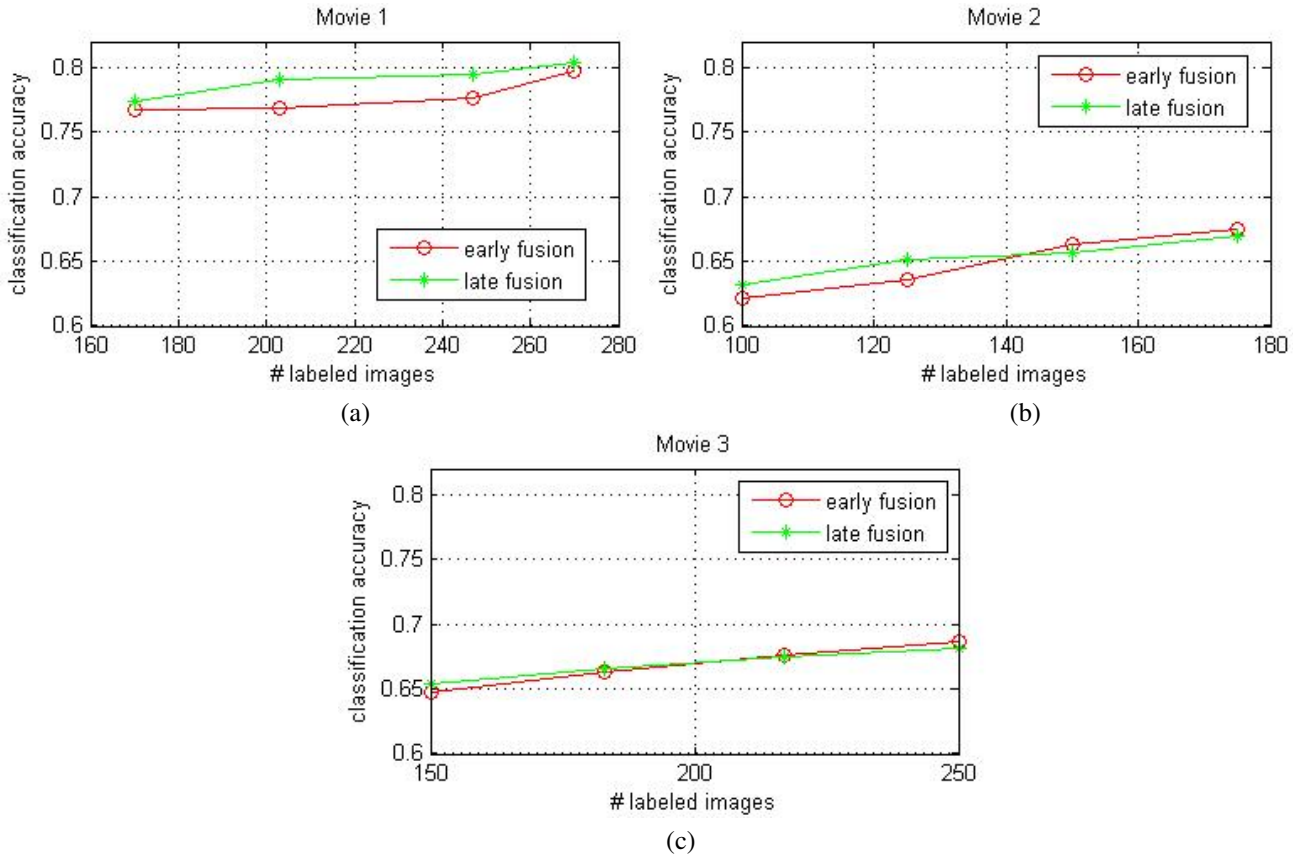


Fig. 3. Classification accuracy of MLPP-PCLP with respect to the number of images that enter the initially labeled data set for (a) Movie 1 (b) Movie 2 and (c) Movie 3.

TABLE V
CLASSIFICATION ACCURACY OF MLPP-CLP WITH EARLY (MLPP-E) AND LATE (MLPP-L) FUSION AND SINGLE-CHANNEL LPP, OLPP, PCLPP, GOLPP AND NPE FOR THREE STEREO MOVIES

| | MLPP-E | MLPP-L | LPP | OLPP | PCLPP | GoLPP | NPE |
|---|---|---|---|---|---|---|---|
| Movie 1 | 79.26% | **80.52%** | 71.39% | 66.47% | 73.21% | 66.34% | 72.75% |
| Movie 2 | **67.10%** | **67.10%** | 53.01% | 46.23% | 57.17% | 53.15% | 54.19% |
| Movie 3 | 68.38% | **68.39%** | 59.30% | 57.74% | 60.25% | 56.00% | 59.48% |
| Average | 71.76% | **72.20%** | 61.67% | 57.46% | 63.83% | 59.32% | 62.54% |

TABLE VI
CLASSIFICATION ACCURACY OF MLPP-CLP WITH EARLY (MLPP-E) AND LATE (MLPP-L) FUSION AND STEREO LPP, OLPP, PCLPP, GOLPP AND NPE FOR THREE STEREO MOVIES

| | MLPP-E | MLPP-L | LPP | OLPP | PCLPP | GoLPP | NPE |
|---|---|---|---|---|---|---|---|
| Movie 1 | 79.26% | **80.52%** | 74.26% | 68.40% | 75.71% | 69.26% | 75.64% |
| Movie 2 | **67.10%** | **67.10%** | 56.05% | 48.68% | 59.77% | 59.52% | 57.46% |
| Movie 3 | 68.38% | **68.39%** | 61.50% | 63.06% | 62.54% | 58.91% | 63.14% |
| Average | 71.76% | **72.20%** | 64.33% | 60.80% | 66.28% | 63.10% | 65.82% |

TABLE VII
CLASSIFICATION ACCURACY OF MLPP-CLP WITH EARLY (MLPP-E) AND LATE (MLPP-L) FUSION AND LPP FOR MODALITIES 1-5 AND MULTI-MODAL LPP.

| | MLPP-E | MLPP-L | LPP-M1 | LPP-M2 | LPP-M3 | LPP-M4 | LPP-M5 | LPP-Multi |
|---|---|---|---|---|---|---|---|---|
| UCF11 1 | 67.71% | **73.01%** | 55.17% | 50.56% | 56.10% | 57.69% | 47.93% | 66.53% |

all data representations. When the value of $\varepsilon$ is large, i.e., the value of the second term in (26) is larger than the value of the first term, then increased significance is given to the minimization of the coefficient vector $\tau$ norm and, thus, $\tau$ takes the values $\tau_k = \frac{1}{K}$, $k = 1, \ldots, K$. This means that, all data representations participate equally in the dimensionality reduction procedure. This result is also undesired, because it does not lead to optimal fusion of the data representations. Thus, the optimal values for $\varepsilon$ that lead to comparable values between the coefficients vector values are the ones that result in comparable values for the two terms of equation (26). In the conducted experiments, such performance was achieved for $\varepsilon = 1$.

## VI. Conclusions

In this paper, a novel method for propagating person identity labels on facial images extracted from stereo videos was introduced. The proposed method operates on data with multiple representations, by calculating a projection matrix that projects the multiple data representation matrices to a reduced dimensionality space that preserves the locality information in the original representations and that satisfies a priori pairwise information in the form of pairwise must-link and cannot-link constraints. Moreover, a novel method was presented for selecting the data set from which label propagation will begin. The initialization method exploits information about the data structure obtained from the application of a simple clustering algorithm. Experimental results on a large data set consisting of facial images extracted from three stereo movies showed that both the subspace representation through MLPP and the label propagation initialization result in an increase in the classification accuracy compared to state of the art methods.

The proposed method finds application in semi-automatic annotation of stereo video content in television broadcaster archives, by enabling the archivists to put person identity labels on the facial images that appear in the videos. Very good facial image semi-automatic annotation results have been obtained by manually labeling only very few facial images in a movie. According to these experiments, we can obtain correct facial identity labels up to $72\%$ of the facial images that appear in the three movies, when the annotator manually labels only $700$ images, that consist only $5\%$ of the facial image data set size and only $0.42\%$ of the total number of facial images in these movies.

## References


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

[2] O. Zoidi, N. Nikolaidis, and I. Pitas, "Exploiting clustering and stereo information in label propagation of facial images," in *IEEE Symposium Series on Computational Intelligence*, 2013.

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

[4] L. Zhang, L. Qiao, and S. Chen, "Graph-optimized locality preserving projections," *Pattern Recognition*, vol. 43, no. 6, pp. 1993–2002, 2010.

[5] J. Lu and Y.-P. Tan, "Regularized locality preserving projections and its extensions for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 3, pp. 958–963, 2010.

[6] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[7] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," *Image and Vision computing*, vol. 24, no. 3, pp. 239–248, 2006.

[8] G.-F. Lu, Z. Lin, and Z. Jin, "Face recognition using discriminant locality preserving projections based on maximum margin criterion," *Pattern Recognition*, vol. 43, no. 10, pp. 3572–3579, 2010.

[9] L. Zhu and S. Zhu, "Face recognition based on orthogonal discriminant locality preserving projections," *Neurocomputing*, vol. 70, no. 79, pp. 1543 – 1546, 2007.

[10] P. T. Pham, T. Tuytelaars, and M.-F. Moens, "Naming people in news videos with label propagation," *IEEE MultiMedia*, vol. 18, no. 3, pp. 44 –55, march 2011.

[11] O. Chapelle, A. Zien *et al.*, *Semi-Supervised Learning*. MIT Press, 2006.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] H. Cevikalp, J. Verbeek, F. Jurie, A. Klaser *et al.*, "Semi-supervised dimensionality reduction using pairwise equivalence constraints," in *3rd International Conference on Computer Vision Theory and Applications (VISAPP'08)*, 2008, pp. 489–496.

[18] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[19] 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.

[20] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2006.

[21] S. Boyd and L. Vanderberghe, *Convex optimization*. Cambridge University Press, 2004.

[22] 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.

[23] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888 –905, Aug 2000.

[24] L. Freeman, "A set of measures of centrality based upon betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

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

[26] 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.

[27] 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.

[28] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1978–1990, 2011.

[29] 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.

[30] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. Int. Conf. Computer Vision (ICCV'05)*, 2005.

[31] 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.

[32] 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.

PLACE PHOTO HERE

**Olga Zoidi** received the B.Sc. in Mathematics in 2004 and the diploma in Electrical and Computer Engineering in 2009, both from the Aristotle University of Thessaloniki. She is currently a researcher and Ph.D. student in the Artificial Intelligence and Information Analysis laboratory at the Department of Informatics of Aristotle University of Thessaloniki. From 2010 to 2013 she was a Teaching Assistant for the lessons Digital Signal Processing and Digital Image processing at the Department of Informatics at the Aristotle University of Thessaloniki. She has co-authored more than 15 papers in international journals and conferences. Her current research interests include image and video processing, computer vision and pattern recognition.

PLACE PHOTO HERE

**Anastasios Tefas (M04)** received the B.Sc. in informatics in 1997 and the Ph.D. degree in informatics in 2002, both from the Aristotle University of Thessaloniki, Greece. Since 2013 he has been an Assistant Professor at the Department of Informatics, Aristotle University of Thessaloniki. From 2008 to 2012, he was a Lecturer at the same University. From 2006 to 2008, he was an Assistant Professor at the Department of Information Management, Technological Institute of Kavala. From 2003 to 2004, he was a temporary lecturer in the Department of Informatics, University of Thessaloniki. From 1997 to 2002, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. Dr. Tefas participated in 12 research projects financed by national and European funds. He has co-authored 39 journal papers, 113 papers in international conferences and contributed 7 chapters to edited books in his area of expertise. Over 2150 citations have been recorded to his publications and his H-index is 23 according to Google scholar. His current research interests include computational intelligence, pattern recognition, statistical machine learning, digital signal and image processing and computer vision, biometrics and security.

PLACE PHOTO HERE

**Nikos Nikolaidis (S92M05-SM09)** received the Diploma of Electrical Engineering and the Ph.D. degree in Electrical Engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1991 and 1997, respectively. From 1992 to 1996, he was a Teaching Assistant at the Departments of Electrical Engineering and Informatics at the Aristotle University of Thessaloniki. From 1998 to 2002, he was a Postdoctoral Researcher and Teaching Assistant at the Department of Informatics, Aristotle University of Thessaloniki, where he is currently an Assistant Professor. He has co-authored 1 book, 15 book chapters, 40 journal papers and 136 conference papers and co-edited one book and two special issues in journals. Moreover he has co-organized 6 special sessions in international conferences. The number of citations to his work by third authors exceeds 3000 (h-index 24). He has participated into 24 research projects funded by the EU and national funds. His areas of interest/expertise include stereoscopic/multiview video processing/analysis, anthropocentric video analysis (human detection and tracking, activity recognition), computer vision, digital image/video processing, computer graphics and visualization, multimedia copyright protection. Dr. Nikolaidis is currently serving as associate editor for the EURASIP Journal on Image and Video Processing, the International Journal of Innovative Computing Information and Control, the Innovative Computing, Information and Control Express Letters and the Journal of Information Hiding and Multimedia Signal Processing. Furthermore, he is a member of the Editorial Board of the International Journal of Multimedia Intelligence and Security. He served as Exhibits chair of IEEE ICIP 2001, Technical Program chair of IEEE IVMSP 2013 workshop and is currently serving as Publicity co-chair of EUSIPCO 2015. He is an IEEE Senior Member.

PLACE PHOTO HERE

**Ioannis Pitas (SM94-F07)** Prof. Ioannis Pitas (IEEE fellow, IEEE Distinguished Lecturer, EURASIP fellow) received the Diploma and PhD degree in Electrical Engineering, both from the Aristotle University of Thessaloniki, Greece. Since 1994, he has been a Professor at the Department of Informatics of the same University. He served as a Visiting Professor at several Universities.

His current interests are in the areas of image/video processing, intelligent digital media, machine learning, human centered interfaces, affective computing, computer vision, 3D imaging and biomedical imaging. He has published over 750 papers, contributed in 39 books in his areas of interest and edited or (co-)authored another 9 books. He has also been an invited speaker and/or member of the program committee of many scientific conferences and workshops. In the past he served as Associate Editor or co-Editor of eight international journals and General or Technical Chair of four international conferences (including ICIP2001). He participated in 68 R&D projects, primarily funded by the European Union and is/was principal investigator/researcher in 40 such projects. He has 17900+ citations (Source Publish and Perish), 6250+ (Scopus) to his work and h-index 64+ (Source Publish and Perish), 38+ (Scopus).