LABEL PROPAGATION ON FACIAL IMAGES USING SIMILARITY AND DISSIMILARITY LABELLING CONSTRAINTS

*Efstratios Kakaletsis, Olga Zoidi, Ioannis Tsingalis, Anastasios Tefas, Nikos Nikolaidis, Ioannis Pitas**

AIIA Lab, Department of Informatics
Aristotle University of Thessaloniki, Greece
Emails:{tefas, nikolaid, pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper, a novel multimedia data (specifically facial images) label propagation method is presented that is based on the inclusion of labelling constraints in the objective function of the MLPP-CLP state of the art algorithm. The proposed method can incorporate pairwise facial image similarity and dissimilarity constraints into the objective function of the aforementioned method. Experiments which have been conducted on facial image labelling in three stereoscopic movies, confirm the increased labelling accuracy of the proposed method.

***Index Terms*—** label propagation, stereoscopic video, facial images

1. INTRODUCTION

Nowadays, semantic annotation (tagging) [1], [2] of large multimedia data archives is typically performed manually by individual users (annotators). Such user-generated image tags can be used in multimedia content search, organization, retrieval and browsing. Tagging facial images in a video sequence is a time consuming task. Prior knowledge, for example, of actor identities and their appearances in specific video frames or shots, can be used to initialize automatic facial image labeling. Indeed, if such information is available, the annotation of the rest of the facial images can be based on label propagation, which spreads labels from a small labeled facial images dataset to a large unlabeled one (Figure 1). Essentially, facial label propagation is a semi-supervised face recognition method.

Different label propagation methods exist. A detailed survey on semi-supervised learning can be found in [3]. Label propagation methods include the use of Gaussian random fields and harmonic functions, as well as the linear neighborhood propagation [4]. Label propagation can be formally described using graphs whose nodes and edges represent the visual data (facial image) representations and their pairwise similarities, respectively [5]. Then, the label inference is an information diffusion process from labeled nodes to unlabeled ones through the graph paths by employing facial image similarities. In iterative label propagation methods, label diffusion is performed gradually on the unlabeled data, according to an update rule [3]. A method for label propagation on similarity graphs can be found in [4]. The proposed algorithm, called Linear Neighborhood Propagation (LNP), can discover the structure of the whole dataset through synthesizing the linear neighborhood around each data object. Another popular label propagation method is the local and global consistency method [6] which is a principled approach to semi-supervised learning that presents a classification function that is sufficiently smooth with respect to the intrinsic structure collectively revealed by known labeled and unlabeled points. Its Multiple-graph Locality Preserving Projections - Cluster-based Label Propagation (MLPP-CLP) variant that operates in data with multiple representations, is described in [7].

In this paper, the main aim of the proposed stereo facial images label propagation method is enhancing labeling accuracy, by incorporating pairwise facial image similarity and dissimilarity constraints into the objective function of the MLPP-CLP algorithm resulting into Constrained MLPP-CLP (CMLPP-CLP). In the proposed approach, positive facial label information of the form 'facial image i belongs to actor j' and negative facial labelling information of the form 'facial image i does not belong to actor j' (labelling constraints) [7] are used. The incorporation of such pairwise similarity and dissimilarity constraints into the objective function of the label propagation leads to an increased face recognition accuracy as proven in the experiments. Label propagation on other types of data can be also performed.

2. MLPP-CLP LABEL PROPAGATION

A short description of the MLPP-CLP approach is presented in this section. The full algorithm can be found in [7], [5]. Assume a set of labeled facial images $X_L = \{\mathbf{x}_i\}_{i=1}^{m_l}$ which have been assigned labels (actor names) from the set $L = \{l_j\}_{j=1}^{Q}$ and a set of unlabeled facial images $X_U = \{\mathbf{x}_i\}_{i=1}^{m_u}$. Their union is given by $X = \{\mathbf{x}_1, ..., \mathbf{x}_{m_l}, \mathbf{x}_{m_l+1}, ..., \mathbf{x}_M\}$, $M = m_l + m_u$ [8]. The objective of label propagation is

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

to spread the facial image labels in L from the set of the labeled images X_L to the set of the unlabeled images X_U, while maintaining local and global labeling consistency [6].

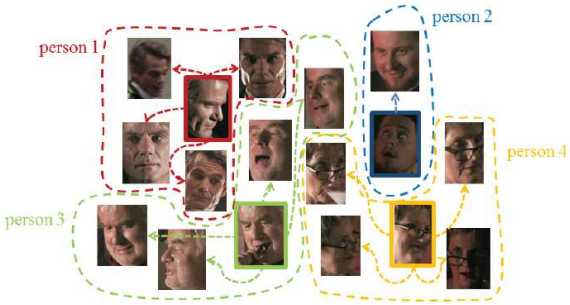

Fig. 1: Label propagation in facial images

The initial information about the labeled data is described by the $M \times Q$ matrix $\mathbf{Y}$:

$$Y_{ij} = \begin{cases} 1, & \text{if image } i \text{ is labeled as } y_i = j \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The algorithm begins with the construction of a symmetric facial image similarity matrix $\mathbf{W} \in \mathbb{R}^{M \times M}$, as described in [5], which represents the facial image similarity graph. The rows/columns of the matrix correspond to the temporally ordered facial images, i.e. the facial images in the sequence they appear in the video. The element W_{ij} of this matrix denotes the similarity between the i-th and the j-th facial image. More specifically, the edge in the graph that connects the nodes (facial images) i and j is assigned with a value W_{ij} that indicates the similarity between these two nodes. This similarity is computed according to the heat kernel equation:

$$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}} \quad (2)$$

where σ is the mean edge length distance among neighbors and x_i is the feature vector used to represent node/image i. The main diagonal $W_{ii}, i = 1, \ldots, M$ of the similarity matrix consists of the similarities of facial images with themselves which is set equal to zero because there is no point to conduct label propagation from a facial image to itself.

Then, vectors $\mathbf{f}_i \in \mathbb{R}^{1 \times Q}, i = 1, ..., M$ that assign a score for every possible actor label to facial image i, thus defining the matrix $\mathbf{F} = [\mathbf{f}_1^T, ..., \mathbf{f}_M^T]^T \in \mathbb{R}^{M \times Q}$, are calculated. More specifically, $\mathbf{F}$ is calculated by minimizing [5]:

$$Q(\mathbf{F}) = \frac{1}{2} tr(\mathbf{F}^T \mathbf{L} \mathbf{F}) + \mu tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})), \quad (3)$$

where $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$ is the normalized facial image similarity graph Laplacian, $\mathbf{D}$ is the diagonal matrix having entries $D_{ii} = \sum_j W_{ij}$ and μ is a regularization parameter. This minimization problem leads to the following solution:

$$\mathbf{F} = (1 - a)(\mathbf{I} - a\mathbf{S})^{-1}\mathbf{Y}, \quad (4)$$

where $a = \frac{1}{1+\mu}$ and $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.

The final facial image label (actor name) is assigned to facial image i according to the following decision rule:

$$y_i = \arg\max_{j \in 1, \ldots, Q} [f_{i1}, \ldots, f_{ij}, \ldots, f_{iQ}]. \quad (5)$$

The regularization framework (3) was extended to the case of label propagation on multiview facial images [7]. In this case, multiple graphs are constructed for the data, one for each one of the K facial image representations (i.e., views, for example two views, $K = 2$, for stereoscopic images). Each graph is represented by the corresponding similarity matrix $\mathbf{W}_k, k = 1, ..., K$. In this case, the regularization framework (3) takes the form:

$$Q(\mathbf{F}, \boldsymbol{\tau}) = \frac{1}{2} \sum_{k=1}^{K} \tau_k tr(\mathbf{F}^T \mathbf{L}_k \mathbf{F}) + \mu tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})), \quad (6)$$

subject to the constraint: $\sum_{k=1}^{K} \tau_k = 1$, that leads to the following optimal solution for $\mathbf{F}$:

$$\mathbf{F} = (1 - a) \left(\mathbf{I} - a \sum_k \tau_k \mathbf{S}_k \right)^{-1} \mathbf{Y}. \quad (7)$$

where τ_k, k = *1,...,K* is the weight that corresponds to the *k*-th data representation and $\mathbf{S}_k = \mathbf{D}^{-1/2}\mathbf{W}_k\mathbf{D}^{-1/2}$.

A method for computing the weights τ_k called Multi-graph Locality Preserving Projections (MLPP) was introduced in [7]. It performs dimensionality reduction [9] of data with multiple representations by constructing a single projection matrix $\mathbf{A}$ for all data representations, while preserving the data locality information in all representations and ensuring additional pairwise similarity and dissimilarity constraints on the data [10]. It should be noted, in our case, these constraints are additionally incorporated in the label propagation objective function. The weights τ_k of each data representation to the construction of the projection matrix $\mathbf{A}$ are the optimal weights for the label propagation cost function (6).

3. CONSTRAINED MLPP-CLP (CMLPP-CLP)

Typically, label propagation techniques assume that facial images (or samples in general), which are similar to each other, should be assigned the same label. In order to increase the facial image recognition accuracy using label propagation, information in the form of pairwise image similarity and dissimilarity constraints can be incorporated. In other words, data that satisfy similarity constraints should be assigned the same label and data that satisfy dissimilarity constraints should be assigned different labels [11]. Let S be the set of similar facial image pairs:

$$S = \{(i, j) | \mathbf{x}_i, \mathbf{x}_j \text{ must have the same label}\} \quad (8)$$

In our case, S contains the facial images that belong to the same facial image trajectory (facial images in consecutive frames derived by a face tracker) and most probably correspond to the same actor. Also let D be the set of dissimilar pairs:

$$D = \{(i,j)|\mathbf{x}_i, \mathbf{x}_j \text{ must have different labels}\} \quad (9)$$

In this paper, facial image pairs that appear on the same frame (and belong to different actors) are included in D.

Two weight matrices $\mathbf{W}_s$, $\mathbf{W}_d$ can be constructed as follows:

$$W_{s,ij} = \begin{cases} 1, & \text{if } (i,j) \in S \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

$$W_{d,ij} = \begin{cases} 1, & \text{if } (i,j) \in D \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

The pairwise constraints can be diffused to neighboring nodes as follows. Let N_i be the neighborhood of the node i, based e.g., on thresholding the Euclidean distance between two nodes $\|\mathbf{x}_i - \mathbf{x}_j\| < e$ and $\mathbf{P} \in \Re^{M \times M}$ be the sparse neighborhood probability matrix:

$$P_{ij} = \begin{cases} \frac{1}{|N_i|} & \text{if } j \in N_i \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

where $|N_i|$ is the cardinality of the set N_i. Pairwise similarity and dissimilarity information is propagated to neighboring nodes according to the iterative procedure:

$$\mathbf{F}_s^{(t)} = \alpha \mathbf{P} \mathbf{F}_S^{(t-1)} + (1-\alpha)\mathbf{W}_s, \quad (13)$$

$$\mathbf{F}_d^{(t)} = \alpha \mathbf{P} \mathbf{F}_D^{(t-1)} + (1-\alpha)\mathbf{W}_d, \quad (14)$$

where the parameter α, $0 \leq \alpha \leq 1$, controls the percentage of information which the node will receive from its neighbors and from the initial state. Iterative equations (13), (14) converge to the steady state solution [12]:

$$\mathbf{F}_s = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{W}_s \quad (15)$$

$$\mathbf{F}_d = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{W}_d. \quad (16)$$

As already stated, the incorporation of the pairwise similarity and dissimilarity constraints into label propagation aims at enhancing face recognition accuracy. This is obtained by the following procedure. First, the pairwise dissimilarity constraints between facial images that appear in the same video frame are taken into account by constructing the new weight matrix $\mathbf{W}_d$ as in (11). Respectively, the pairwise similarity constraints between the facial images that appear in the same facial image trajectory are taken into account by constructing the new weight matrix $\mathbf{W}_s$ as in (10). Then, the MLPP method, briefly described in Section II, is carried out, in order to perform data dimensionality reduction by preserving the pairwise similarity and dissimilarity information. Subsequently, label propagation is performed on the data projections, by incorporating the pairwise similarity and dissimilarity constraints to the objective function of label propagation, as follows:

$$Q(\mathbf{F}) = \frac{1}{2} tr(\mathbf{F}^T \left(\sum_{k=1}^{K} \tau_k \mathbf{L_k} + \beta \mathbf{L_s} - \gamma \mathbf{L_d} \right) \mathbf{F}) + \mu tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})) \quad (17)$$

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 constrains, respectively. $\mathbf{L}_k$ varies according to the data representation, while $\mathbf{L}_s$, $\mathbf{L}_d$ are constant for all representations. Finally, $\mathbf{D}_s, \mathbf{D}_d$ and $\mathbf{D}_k$ are the diagonal degree matrices with entries $D_{s,ii} = \sum_{j=1}^{M} F_{s,ij}$, $D_{d,ii} = \sum_{j=1}^{M} F_{d,ij}$ and $D_{k,ii} = \sum_{j=1}^{M} W_{k,ij}$. The parameters β, γ are chosen in such a way that the matrix $\sum_{k=1}^{K} \tau_k \mathbf{L_k} + \beta \mathbf{L_s} - \gamma \mathbf{L_d}$ is positive definite. Minimization of equation (17) leads to the following label propagation solution:

$$\mathbf{F} = \mu \left(a\mathbf{I} + \sum_{k=1}^{K} \tau_k \mathbf{L}_k + \beta \mathbf{L}_s - \gamma \mathbf{L}_d \right)^{-1} \mathbf{Y}. \quad (18)$$

4. EXPERIMENTS

Experimental evaluation of the proposed technique was performed on facial image label propagation in three stereoscopic movies. Person identity (label) propagation was performed on the facial images that appear in the two (left, right) video channels of these movies. First, dimensionality reduction is applied to the facial image regions of interest (ROI) in each channel separately according to MLPP method. The data dimensionality is reduced from 1271 (for a facial image region of size 41×31 pixels) to 75 dimensions. Then, label propagation is performed. For label propagation initialization, K-means clustering was used and only 5% of the facial images are manually labeled. As we have two ($K = 2$) different data representations on stereo video, namely the left and right stereo channels, late fusion [7] of the two data representations was performed. Both similarity and dissimilarity constraints were incorporated in the objective function of label propagation. Since the total number of the extracted facial images using the face detector [13] and the face tracker [14] is very large, experiments have been conducted with a small dataset of facial images from every movie. In total, 13850 images were selected from the three movies. More precisely, each facial image trajectory is represented by one image (short trajectories) or more images (longer trajectories).

Experimental results of the Constrained MLPP-CLP (CMLPP-CLP), on the three stereo movies, are shown in Table I in comparison to those of MLPP-CLP. Similarity constraints involve facial images in the same facial trajectory. All pairs of such images are included in set S in (8) since we assume that they depict the same person. Dissimilarity constraints involve facial images in the same frame. All pairs of such images are included in set D in (9) since they obviously depict different persons. We notice that incorporation of the pairwise constraints into the objective function of label

propagation increases the classification accuracy on average by approximately 4.6%. It should be stressed here that, as detailed in [7], MLPP-CLP outperforms other state of the art label propagation techniques, such as those in [15–19]. Thus, by showing, that the proposed CMLPP-CLP approach performs better than MLPP-CLP, we also show that it performs better than the other state of the art approaches.

5. CHOICE OF "CONSTRAINTS VS LABELS" STRATEGY

In an incremental semi-automatic label propagation task, the classification accuracy obtained, when a certain percentage of facial images is labeled, may be unsatisfactory. Then, a human annotator can perform two different actions towards reaching a desired classification accuracy: a) he can manually label additional unlabeled images or b) he can place additional pairwise facial image similarity or dissimilarity constraints. Thus the following questions naturally arise with respect to a "constraints vs labels" strategy: which of the two actions is more beneficial? What is the effect, in a certain label propagation problem of a) inserting one more constraint or b) labeling one more unlabeled image? Investigations towards answering these questions were conducted. Specifically, given a desired classification accuracy, the ratio of additional images that have to be manually labeled in order to achieve this accuracy, versus the required number of pairwise constraints needed to obtain the same accuracy was calculated. More specifically, let N_{RL} be the current number of manually labeled images, N_{TL} be the number of manually labeled images required in order to reach the desired classification accuracy P (without the use of any pairwise constraints) and N_c be the number of pairwise image similarity constraints needed (in addition to the N_{RL} labeled images) in order to reach P. For a given value of N_{TL} (and thus for a certain desired classification accuracy) and for a certain number of N_{RL}, the ratio r of the additional labeled images needed to reach P over the required pairwise constraints N_c needed to achieve the same target accuracy is given by: $r = (N_{TL} - N_{RL})/N_c$. Large values of r , e.g. values close to 1, mean that adding a labeling constraint has almost the same effect on the classification accuracy as labeling one more unlabeled image. On the contrary, small values of r denote that much more additional facial image constraints than labels are needed in order to reach the desired accuracy P. For example, $r = 0.5$ means that, in order to reach P, one needs to place twice as many labeling constraints than additional labels on images in order to reach P. We have conducted such an experiment in one video, having 116 facial images. Plots of r versus N_{RL} for various values of N_{TL} are shown in Figure 2. Six curves are depicted in the Figure 2, each corresponds to a fixed N_{TL} value, which in turn corresponds to a certain classification accuracy value P. This figure suggests that r increases in general, as N_{RL} increases and never goes above 1. This means that, as the number of labeled images increases, the effect of the constraints increases. However, since r is in more cases significantly below 1, always less additional labeled images are needed than additional labeling constraints to reach the desired accuracy. The above investigation naturally involves not only the effect of a "labels vs constraints" strategy towards reaching a certain (target) classification accuracy but also the cost/effort associated with each of the two actions (adding a label or a constraint). Usually, the effort of labeling an image is larger than that of assigning a pairwise labeling constraint.

Table 1: Face recognition accuracy obtained using MLPP-CLP or Constrained MLPP-CLP (CMLPP-CLP) label propagation methods.

	MLPP-CLP	CMLPP-CLP
Movie 1	0.7859	0.801223
Movie 2	0.6395	0.672213
Movie 3	0.62	0.710133

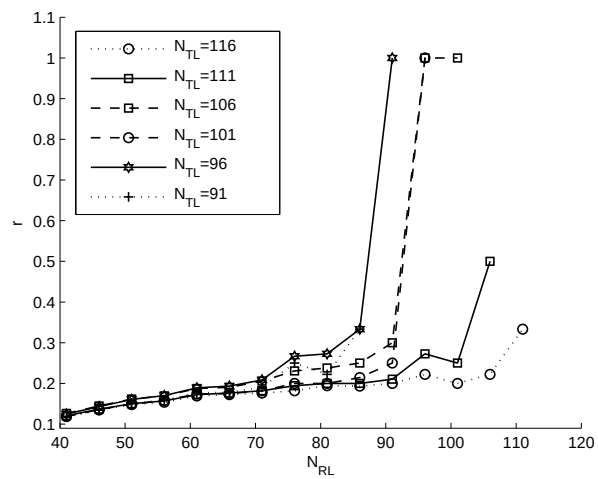

Fig. 2: Ratio of the number additional labeled images over the number of additional labeling constraints for achieving desired facial image classification accuracy.

6. CONCLUSIONS

In this paper, a novel method for propagating person identity labels on facial images extracted from stereo videos was introduced. The proposed method which operates on data with multiple representations (but can easily be adopted to work on single representation data and data different than facial images) aims at enhancing the state of the art MLPP-CLP label propagation method [7]. Experiments show that the use of similarity and dissimilarity labelling constraints in the objective function of label propagation increase the classification accuracy.

7. REFERENCES

[1] L Ballan, M Bertini, A Del Bimbo, and G Serra, "Enriching and localizing semantic tags in internet videos," *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1541–1544, 2011.

[2] Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti, and Yiannis Kompatsiaris, "A survey of semantic image and video annotation tools," *In: Paliouras G., Spyropoulos C.D., Tsatsaronis G. (eds) Knowledge-driven multimedia information extraction and ontology evolution, Lecture Notes in Computer Science*, vol. 6050, pp. 196–239, Springer, Berlin, Heidelberg, 2011.

[3] X Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005, http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

[4] F Wang and C Zhang, "Label propagation through linear neighborhoods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 55–67, 2008.

[5] O Zoidi, A Tefas, N Nikolaidis, and I Pitas, "Iterative label propagation on facial images," *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, pp. 1222–1226, 2014.

[6] D Zhou, O Bousquet, T N Lal, J Weston, and B Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, pp. 321–328, 2004.

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

[8] O Zoidi, E Fotiadou, N Nikolaidis, and I Pitas, "Graph-based label propagation in digital media: A review," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 48, 2015.

[9] L JP van der Maaten, E O Postma, and H J van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.

[10] G Yu, H Peng, J Wei, and Q Ma, "Robust locality preserving projections with pairwise constraints," *Journal of Computational Information Systems*, vol. 6, no. 5, pp. 1631–1636, 2010.

[11] O Zoidi, N Nikolaidis, and I Pitas, "Semi-supervised dimensionality reduction on data with multiple representations for label propagation on facial images," *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6019–6023, 2014.

[12] G H Golub and C F Van Loan, "Matrix computations," vol. 3, 2012.

[13] GN Stamou, M Krinidis, N Nikolaidis, and I Pitas, "A monocular system for automatic face detection and tracking," *Visual Communications and Image Processing 2005*, vol. 5960, pp. 794–802, 2005.

[14] O Zoidi, A Tefas, and I Pitas, "Visual object tracking based on local steering kernels and color histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870–882, 2013.

[15] X He and P Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[16] L Zhu and S Zhu, "Face recognition based on orthogonal discriminant locality preserving projections," *Neurocomputing*, vol. 70, no. 7-9, pp. 1543–1546, 2007.

[17] H Cevikalp, J Verbeek, F Jurie, and A Klaser, "Semi-supervised dimensionality reduction using pairwise equivalence constraints," in *VISAPP'08 - 3rd International Conference on Computer Vision Theory and Applications*. INSTICC, 2008, vol. 1, pp. 489–496.

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

[19] X He, D Cai, S Yan, and H-J Zhang, "Neighborhood preserving embedding," in *10th IEEE International Conference on Computer Vision, 2005. ICCV 2005*. IEEE, 2005, vol. 2, pp. 1208–1213.