Kakaletsis, E., Zoidi, O., Tefas, A., Nikolaidis, N., & Pitas, I. (2016). Fast Label Propagation on Facial Images Using a Pruned Similarity Matrix. In Digital Media Industry & Academic Forum (DMIAF). Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/DMIAF.2016.7574911

Fast Label Propagation on Facial Images Using a Pruned Similarity Matrix

Efstratios Kakaletsis*, Olga Zoidi*, Anastasios Tefas*, Nikos Nikolaidis*, Ioannis Pitas*†
*Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
†Department of Electrical and Electronic Engineering, University of Bristol, UK
{tefas, nikolaid}@aiia.csd.auth.gr

***Abstract*—The label propagation process, which is often used to semantically annotate (tag) large amounts of multimedia data assets must be fast, in order to be efficient. In this paper, a novel facial images fast labeling method that is essentially a semi-supervised face recognition approach, is presented. The proposed method is based on the acceleration of a state of the art facial identity label propagation technique. The new method is called pruned label propagation due to the fact that the facial label inference is conducted using a similarity matrix containing fewer entries, namely the pairwise similarities that reside in the main and the off-diagonals of this matrix. Experiments conducted on facial image labeling in three stereoscopic movies, confirm the increased labeling accuracy and the reduced computational complexity of the proposed method.**

I. Introduction

Nowadays, the semantic annotation (tagging) [1], [2] of large multimedia data archives is typically performed manually by individual users (annotators). More specifically, the semantic annotation of facial images with the names of the depicted people is popular in social media sites, such as Flickr [3], and Facebook (for images), or YouTube, Vimeo (for videos). Such user-generated, facial image tags can be used in multimedia content search, retrieval and browsing. Tagging facial images in a video sequence is a time consuming task. Manual labeling of people and their appearances in specific video frames or shots can be used to initialize facial image labeling. Then, the annotation of the rest of facial images can be based on label propagation, which spreads labels from a small labeled facial image dataset to a large unlabeled one. Label propagation can be typically used when full manual annotation is prohibitively slow and/or expensive. Essentially, facial label propagation is a semi-supervised face recognition method. In this paper, we propose an approach for the speedup of the state of the art Multiple-graph Locality Preserving Projections - Cluster-based Label Propagation (MLPP-CLP) method [4] by performing approximate label propagation using a pruned facial image similarity matrix.

Due to this fact, the proposed method is called pruned label propagation. More specifically, instead of the full facial image similarity matrix, only its main diagonal and some off-diagonal entries are used, by exploiting the available temporal ordering of facial images. The facial images are extracted by performing automatic face detection and tracking in the two views of a stereo video [4], resulting in the so called facial image trajectories. Facial image trajectories consist of regions of interest (ROIs) representing detected facial images of size $N_x \times N_y$ pixels. Since the facial images resulting from the face detector and the tracker are temporally ordered, image similarities are calculated only for the temporally nearest neighbors and are stored in a band around the main diagonal of the similarity matrix.

The rest of this paper is organized as follows: Section II provides an overview of the state of the art MLPP-CLP label propagation method [4]. Section III describes the details of the proposed method. In Section IV we present the facial labeling dataset and the experiments which have been conducted to measure the facial recognition accuracy and the reduction in computational complexity. Finally, conclusions are presented in Section V.

II. MLPP-CLP facial image label propagation

A very short description of the MLPP-CLP approach is presented in this section. The full algorithm can be found in [4], [5].

Given a set of labeled facial images $X_L = \{\mathbf{x}_i\}_{i=1}^{m_l}$ which are assigned labels (actor names) from the set $L = \{l_j\}_{j=1}^{Q}$ and a set of unlabeled data $X_U = \{\mathbf{x}_i\}_{i=1}^{m_u}$, their union is given by $X = \{\mathbf{x}_1, ..., \mathbf{x}_{m_l}, \mathbf{x}_{m_l+1}, ..., \mathbf{x}_M\}, M = m_l + m_u$ [6]. The objective of label propagation is 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 [7]. The information about the initially (e.g. manually) labeled data is described by the $M \times Q$ matrix $\mathbf{Y}$, defined as:

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

The algorithm begins with the construction of a facial image similarity matrix $\mathbf{W}$, as described in [5], which represents the facial image similarity graph. 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 nodes similarity. 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. The construction of such a matrix has computational complexity and memory requirements of the order $O(M^2)$ even if an k nearest neighbor (NN) matrix [4] is constructed.

The algorithm utilizes vectors $\mathbf{f}_i, i = 1, ..., M$ that assign a score for every possible actor label to facial image i, thus defining the matrix $\mathbf{F} = [\mathbf{f}_1^T, ..., \mathbf{f}_M^T]^T \in \mathbb{R}^{M \times Q}$ that 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:

$$\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}, \quad (5)$$

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

$$y_i = \arg\max_{j \in 1,...,Q} F_{ij}. \quad (6)$$

The regularization framework (3) can be easily extended to the case of label propagation on multiview facial images. In this case, multiple graphs are constructed for the data, one for each one of the K facial image representations (e.g., views) each of these graphs is represented by the correspoding similarity matrix $W_k, k = 1..K$. In this case, the regularization framework (3) takes the form:

$$Q(\mathbf{F}, \boldsymbol{\tau}) = \frac{1}{2} \sum_{k=1}^{K} \tau_k tr(\mathbf{F}^T \mathbf{L}_k \mathbf{F}) + \mu tr((\mathbf{F} - \mathbf{Y})^T (\mathbf{F} - \mathbf{Y})), \quad (7)$$

subject to the constraint:

$$\sum_{k=1}^{K} \tau_k = 1, \quad (8)$$

that leads to the optimal solution for $\mathbf{F}$:

$$\mathbf{F} = (1 - a) \left(\mathbf{I} - a \sum_k \tau_k \mathbf{S}_k \right)^{-1} \mathbf{Y}. \quad (9)$$

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 [4]. It performs dimensionality reduction [8] 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 [9]. The weight 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 (7), given that the data feature extraction was performed according to MLPP.

III. Proposed Method

A. Pruned Label Propagation

The proposed novel label propagation facial image technique employs a pruned facial image similarity matrix $\mathbf{W}$. Despite the fact that the proposed technique is a well known method in mathematics (utilizing the band matrix for accelerating the solution of a linear system), the temporal order of the facial images in the facial image trajectories which are derived from face detection and especially from face tracking in consecutive frames, is exploited here. More specifically, 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. We assume that all images in a facial image trajectory correspond to the same person and thus for the label propagation we use only the first image of each trajectory. The remaining images in each facial image trajectory adopt the label assigned to the first image of the trajectory by the label propagation procedure. The similarities of the utilized images (in the main diagonal) and the temporally nearest neighbours (in off-diagonals) are stored in the band of the similarity matrix. To this end, the proposed method is accomplished using the following approach.

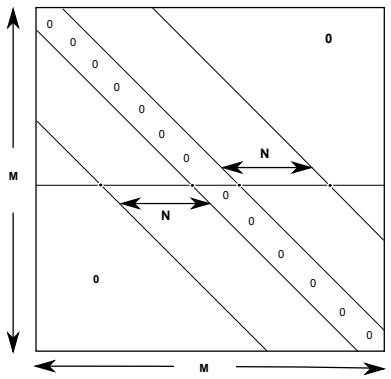

Fig. 1: $(2N + 1)$-band similarity matrix

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. Around these diagonal elements, we calculate only the entries of the N upper and lower diagonals that contain the similarities of the temporally nearest neighbouring facial images, as shown in Figure 1. The similarity matrix is computed according to the Gaussian heat kernel equation:

$$W_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}}, & i \neq j,\ x_i, x_j \text{ are k-NN} \\ & \in N \text{ upper/lower diagonals} \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where $\mathbf{x}_i, \mathbf{x}_j \in \Re^M$ are the feature vectors of the i-th and j-th facial images and σ is a diffusion parameter. Obviously,

$W_{ij} = W_{ji}$. Furthermore, for a band similarity matrix of the form (10), $\mathbf{S}$ (5) is a band matrix as well.

The construction of a band similarity matrix has computational complexity $O(2NM) \simeq O(NM)$, which is much less than the computational complexity $O(M^2)$ of constructing a full $M \times M$ k nearest neighbors (NN) similarity matrix, since $N \ll M$. The experiments have shown that the classification accuracy using either the full (namely the k-NN similarity matrix) or the band similarity matrix for label propagation is approximately the same.

B. Computational Complexity Study

The creation of the matrix $\mathbf{S}$ according to (5) has complexity $O(M^2)$ due to multiplication of the full matrix ($\mathbf{W}$) with diagonal matrices ($\mathbf{D}^{-1/2}$). Moreover, the label propagation solution (4) employing matrix ($\mathbf{I} - a\mathbf{S}$) inversion has complexity $O(M^3)$ [10] and multiplication with the matrix $\mathbf{Y}$ has complexity $O(M^2Q)$. However, solution of (4) by inverting a band matrix using connectivity of Schur's complements [11] has complexity $O(M^2N) + O(M^2Q)$. The first term refers to the inversion of the matrix ($\mathbf{I} - a\mathbf{S}$) whereas the second one refers to the multiplication with matrix $\mathbf{Y}$.

Thus, we can conclude that the computational complexity of the proposed approach including the label propagation procedure is $O(M^2 + M^2N + M^2Q + NM) \simeq O(M^2)$ which is much smaller than the complexity of label propagation involving the full similarity matrix $O(M^2 + M^3 + M^2Q + M^2) \simeq O(M^3)$.

IV. Experiments

A. Dataset and Method Application

Experimental evaluation of the proposed technique was performed on facial image label propagation in three stereoscopic full length movies having total duration 6 hours, 4 minutes and 16 seconds and 546,400 frames in total. Person identity (label) propagation was performed on the facial images that appear in the two (left, right) video channels of these movies. Firstly, dimensionality reduction according to MLPP [4] method, is applied to the facial image regions of interest (ROI) in each channel separately. The data dimensionality is reduced from 1271 (for a facial image region of size 41×31 pixels) to 75 dimensions. Then, the label propagation is performed. For label propagation initialization, the method in [4], that involves K-means clustering, was used and only 5% of the facial images were manually labeled. As we have two ($K = 2$) different data representations on stereo video namely the left and right stereo channels, late fusion [4] of two data representations was performed. The band similarity matrix was computed according to (10).

From the facial images extracted using the face detector [12] and the single channel face tracker [13] only the first image from each facial trajectory has been used in the dataset. In total, 13850 images were used from the three movies, namely 5398, 3498 and 4954 facial images, respectively.

B. Pruned Label Propagation Performance

In this section, we examine the effect of similarity matrix pruning on label propagation, measured by the obtained face recognition accuracy.

Figure 2 displays face recognition accuracy versus the percentage of the retained entries of the full similarity matrix $\alpha_p = \frac{2NM}{M^2} = \frac{2N}{M}$. The horizontal lines show the classification accuracy of the full similarity matrix MLPP-CLP [4] method, which does not depend on a_p. This figure shows that the classification accuracy of the pruned label propagation in one of the three movies (Movie 3) outperforms the classical MLPP-CLP method for most values of a_p most probably due to the fact that the similarity matrix pruning removes noise (semantically-unrelated facial images) from the graph which represents the similarity matrix. For the other two movies the proposed approach has almost equal (Movie 2) or similar but inferior performance (Movie 1) to the MLPP-CLP method, for certain values of a_p. Moreover, we can notice that the classification accuracy for one movie (Movie 1) increases as the percentage a_p increases. However, in Movies 2, 3, the classification accuracy decreases after a value of the a_p (namely $a_p = 0.15$ in Movie 2 and $a_p = 0.2$ in Movie 3). This can be attributed to the introduction of similarity matrix entries which offer additional useful information until a point (value of a_p). After this point, the additional entries correspond to noise and, as a result, the classification accuracy decreases.

Regarding computational complexity, let T_f, T_p be the execution time for the calculation of the full and the band similarity matrix, respectively for the three movies. Figure 3 shows plots of the ratio $r_1 = \frac{T_f}{T_p}$ versus the percentage of the retained entries around the main diagonal (a_p) for the three movies. As the r_1 is always bigger than one, pruning accelerates similarity matrix construction. Moreover, r_1 decreases towards one as a_p increases which is expected since as a_p tends to one, the pruned similarity matrix tends to the full matrix. As can be also observed in Figure 3 the computational savings for a_p values that provide best classification accuracy results, e.g. $a_p = 0.15, 0.2$ are significant (almost 6.94 and 5.36 times faster respectively). The plots follow very well the theoretical relation between r_1 and a_p which is $r_1 = \frac{T_f}{T_p} = \frac{M^2}{2MN} = \frac{M}{2N} = \frac{1}{a_p}$.

Moreover, let T_{LP_f}, T_{LP_p} be the execution time of the label propagation procedure involving the full and the pruned similarity matrix as described in (4). Figure 4 shows the ratio $r_2 = \frac{T_{LP_f} + T_f}{T_{LP_p} + T_p}$, of the total execution times $T_{LP_f} + T_f$ and $T_{LP_p} + T_p$ versus the percentage of retained entries a_p. One can notice that the total matrix construction and label propagation execution time is considerably smaller for the proposed pruning method. The noticed speedup of the whole label propagation procedure including the construction time of the similarity matrix compared to the full similarity matrix approach is for example almost 4.07-3.84 in Movie 3 for values $a_p = 0.15, 0.2$ in which the best recognition accuracy is presented.

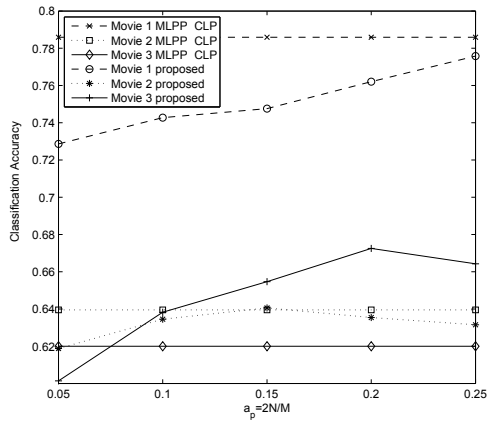

Fig. 2: Face recognition accuracy vs approximation percentage (percentage of retained similarity matrix entries).

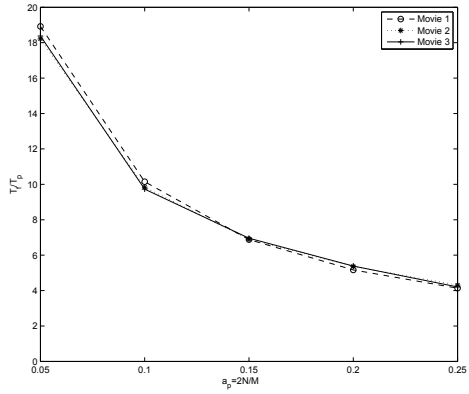

Fig. 3: Ratio of similarity matrix construction time between method [4] and proposed pruning method vs percentage a_p of the retained image similarity entries

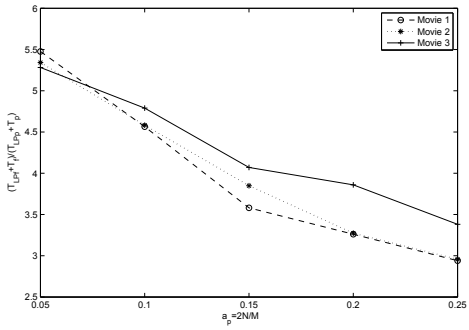

Fig. 4: Ratio of similarity matrix construction time and label propagation execution time between method [4] and proposed pruning method vs percentage a_p of the retained image similarity entries

V. 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 multimedia data with multiple representations, is called pruned label propagation and acts as a fast facial image labeling method. Experiments on a data set consisting of facial images extracted from three stereo movies show that a significant speedup is obtained by creating a band similarity matrix, which contains fewer pairwise facial image similarities. Such a speedup is also achieved in many cases by an increase in the recognition accuracy as the similarity matrix pruning acts as denoising filter upon this matrix.

VI. Acknowledgment

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.

References

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

[2] S Dasiopoulou, E Giannakidou, G Litos, P Malasioti, and Y Kompatsiaris, "A survey of semantic image and video annotation tools," *Knowledge-driven multimedia information extraction and ontology evolution*, pp. 196–239, 2011.

[3] P Andrews, S Kanshin, J Pane, and I Zaihrayeu, "Semantic annotation of images on flickr," *8th European Semantic Web Conference*, pp. 476–480, 2011.

[4] 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.

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

[6] 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.

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

[8] Laurens JP Van der Maaten, Eric O Postma, and H Jaap Van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.

[9] 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.

[10] J. B. Fraleigh and R. A. Beauregard, "Linear algebra," 1987.

[11] A Mahmood, DJ Lynch, and LD Philipp, "A fast banded matrix inversion using connectivity of schur's complements," *IEEE International Conference on Systems Engineering*, pp. 303–306, 1993.

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

[13] 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.