ABSTRACT

Title of dissertation: FACE RECOGNITION FROM WEAKLY LABELED DATA

Ching-Hui Chen, Doctor of Philosophy, 2016

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

Recognizing the identity of a face or a person in the media usually requires lots of training data to design robust classifiers, which demands a great amount of human effort for annotation. Alternatively, the weakly labeled data is publicly available, but the labels can be ambiguous or noisy. For instance, names in the caption of a news photo provide possible candidates for faces appearing in the image. Names in the screenplays are only weakly associated with faces in the videos. Since weakly labeled data is not explicitly labeled by humans, robust learning methods that use weakly labeled data should suppress the impact of noisy instances or automatically resolve the ambiguities in noisy labels.

We propose a method for character identification in a TV-series. The proposed method uses automatically extracted labels by associating the faces with names in the transcripts. Such weakly labeled data often has erroneous labels resulting from errors in detecting a face and synchronization. Our approach achieves robustness to noisy labeling by utilizing several features. We construct track nodes from face and person tracks and utilize information from facial and clothing appearances.

We discover the video structure for effective inference by constructing a minimum-distance spanning tree (MST) from the track nodes. Hence, track nodes of similar appearance become adjacent to each other and are likely to have the same identity. The non-local cost aggregation step thus serves as a noise suppression step to reliably recognize the identity of the characters in the video.

Another type of weakly labeled data results from labeling ambiguities. In other words, a training sample can have more than one label, and typically one of the labels is the true label. For instance, a news photo is usually accompanied by the captions, and the names provided in the captions can be used as the candidate labels for the faces appearing in the photo. Learning an effective subject classifier from the ambiguously labeled data is called ambiguously labeled learning. We propose a matrix completion framework for predicting the actual labels from the ambiguously labeled instances, and a standard supervised classifier that subsequently learns from the disambiguated labels to classify new data. We generalize this matrix completion framework to handle the issue of labeling imbalance that avoids domination by dominant labels. Besides, an iterative candidate elimination step is integrated with the proposed approach to improve the ambiguity resolution.

Recently, video-based face recognition techniques have received significant attention since faces in a video provide diverse exemplars for constructing a robust representation of the target (i.e., subject of interest). Nevertheless, the target face in the video is usually annotated with minimum human effort (i.e., a single bounding box in a video frame). Although face tracking techniques can be utilized to associate faces in a single video shot, it is ineffective for associating faces across multiple

video shots. To fully utilize faces of a target in multiples-shot videos, we propose a target face association (TFA) method to obtain a set of images of the target face, and these associated images are then utilized to construct a robust representation of the target for improving the performance of video-based face recognition task.

One of the most important applications of video-based face recognition is outdoor video surveillance using a camera network. Face recognition in outdoor environment is a challenging task due to illumination changes, pose variations, and occlusions. We present the taxonomy of camera networks and discuss several techniques for continuous tracking of faces acquired by an outdoor camera network as well as a face matching algorithm. Finally, we demonstrate the real-time video surveillance system using pan-tilt-zoom (PTZ) cameras to perform pedestrian tracking, localization, face detection, and face recognition.

FACE RECOGNITION FROM WEAKLY LABELED DATA

by

Ching-Hui Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Professor Behtash Babadi
Professor Larry S. Davis
Professor David W. Jacobs
Professor Min Wu

Dedication

To my family.

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor, Prof. Rama Chellappa, for his guidance and support over the years. His insight on research and his dedication to work have been inspiration to me during my doctoral studies.

Besides, I would like to thank the committee members in my dissertation examination, Prof. Behtash Babadi, Prof. Larry S. Davis, Prof. David W. Jacobs, Prof. Min Wu, for providing valuable comments during my presentation.

I would also like to thank Prof. Vishal M. Patel for his encouragement and fruitful discussions at different stages of my research.

I would also like to thank Jun-Cheng Chen and Dr. Garrett Warnell for sharing their knowledge and experience on controlling the pan-tilt-zoom cameras. Besides, I would like to thank Fritz McCall and Derek Yarnell for the technical support of the camera networks. Without their support, I would not be able to successfully implement the real-time video surveillance system.

I would also like to thank the support of the JANUS team in our group. I would like to thank Carlos D. Castillo, Rajeev Ranjan, Swaminathan Sankaranarayanan, Jun-Cheng Chen for the collaborations on the research project.

Moreover, I would also like to thank Dr. Yi-Chen Chen, Dr. Qiang Qiu, Dr. Ming Du, Dr. Jie Ni, Dr. Jingjing Zheng and Hongyu Xu for teaching me several techniques in the computer vision. I would also like to thank Janice Perrone, Melanie Prange and Arlene Schenk for the administrative help.

Last but not the least, I would like to thank my family.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In this dissertation, we discuss several methods for recognizing the identity of faces and persons in still images or videos. The objective is to accomplish the identification task using weakly labeled data. As compared to human annotation, which requires lots of human effort, the weakly labeled data is usually publicly available but the labeling can be noisy. An effective approach should suppress the impact of noisy samples and resolve the ambiguities in weakly labeled data. Moreover, the face images in videos usually have various poses, which contribute to diverse data, useful for learning a robust classifier. To leverage this advantage, face images of the subject of interest appearing in the scene should be consistently associated, and the associated face images enable us to to utilize the diverse information in the video.

In this chapter we briefly describe these topics.

1.1 Character Identification in TV-series via Non-local Cost Aggregation

We propose a non-local cost aggregation algorithm to recognize the identity of face and person tracks in a TV-series. In our approach, the fundamental element for identification is a track node, which is built on top of face and person tracks. Track

nodes with temporal dependency are grouped into a knot. These knots then serve as basic units in the construction of a k-knot graph for exploring the video structure. We build the minimum-distance spanning tree (MST) from the k-knot graph such that track nodes of similar appearance are adjacent to each other in MST. Non-local cost aggregation is performed on MST, which ensures that information from face and person tracks is utilized as a whole to improve the identification performance. The identification task is performed by minimizing the cost of each knot, which takes into account the unique presence of a subject in a venue. Experimental results demonstrate the effectiveness of our method.

1.2 Learning from Ambiguously Labeled Face Images

Learning a classifier from ambiguously labeled face images is challenging since training images are not always explicitly-labeled. For instance, face images of two persons in a news photo are not explicitly labeled by their names in the caption. We propose a Matrix Completion for Ambiguity Resolution (MCar) method for predicting the actual labels from ambiguously labeled images. This step is followed by learning a standard supervised classifier from the disambiguated labels to classify new images. To prevent the majority labels from dominating the result of MCar, we generalize MCar to a weighted MCar (WMCar) that handles label imbalance. Since WMCar outputs a soft labeling vector of reduced ambiguity for each instance, we can iteratively refine it by feeding it as the input to WMCar. Nevertheless, such an iterative implementation can be affected by the noisy soft labeling vectors, and thus

the performance may degrade. The proposed Iterative Candidate Elimination (ICE) procedure makes the iterative ambiguity resolution possible by gradually eliminating a portion of least likely candidates in ambiguously labeled faces. We further extend MCar to incorporate the labeling constraints between instances when such prior knowledge is available. Compared to existing methods, our approach demonstrates improvements on several ambiguously labeled datasets.

1.3 Video-Based Face Association and Identification

In this work, we focus on a new video-based face identification task, where the target (i.e., person of interest) in the probe video is only annotated once with a face bounding box in a frame and the video may consist of multiple shots. Most of the video face identification techniques assume that the video is of single shot, and thus frame-by-frame bounding boxes of the target face can be extracted by tracking a face across the video frames. Nevertheless, such automatic annotation is vulnerable to the drifting of the face tracker, and the face tracking algorithm is inadequate to associate the face images of the target across multiple shots. We propose a target face association (TFA) technique that retrieves a set of representative face images in a given video that are likely to have the same identity as the target face. These face images are then utilized to construct a robust face representation of the target face for searching the corresponding subject in the gallery. Since two faces that appear in the same video frame cannot belong to the same person, such cannot-link constraints are utilized for learning a target-specific linear classifier for

establishing the intra/inter-shot face association of the target. Experimental results on the newly released JANUS challenge set 3 (JANUS CS3) dataset show that the proposed method generates robust representations from target-annotated videos and demonstrates good performance for the task of video-based face identification problem.

1.4 Face Recognition Using an Outdoor Camera Network

Face recognition in outdoor environments is a challenging task due to illumination changes, pose variations, and occlusions. We discuss several techniques for continuous tracking of faces acquired by an outdoor camera network as well as a face matching algorithm. Active camera networks are capable of reconfiguring the camera parameters to collaboratively capture the close-up views of face images. Robust face recognition methods can utilize compact representations extracted from multi-view videos. Constraints such as consistent tracking of faces and the limitations of network resources should be satisfied. Lastly, we discuss some remaining challenges and emerging frameworks for face recognition in outdoor camera networks.

1.5 Contributions of the Dissertation

We make the following contributions in this dissertation.

- Character identification in TV-series

    1. We propose a unified approach for identifying the face and person tracks in a TV-series by constructing the track nodes to multiplex the modalities

of face and clothing feature from face and person tracks, respectively.

2. We explore the video structure via constructing the minimum-distance spanning tree (MST) from the track nodes such that track nodes that are likely to have the same identity are adjacent to each other. The non-local cost aggregation method is effective in predicting the identity of track nodes.

- Learning from ambiguously labeled face images

1. We propose a matrix completion for ambiguity resolution (MCar) method, where instances and their associated ambiguous labels are jointly considered for disambiguating the class labels.

2. We provide a geometric interpretation of the matrix completion framework from the perspective of recovering the potentially-separable convex hulls of each class.

3. We expand MCar to resolve the label ambiguity in the presence of labeling imbalance.

4. We propose the ICE approach to improve the reliability of iterative WMCar. The integration of WMCar and ICE is effective for resolving the ambiguity and outperforms WMCar in general.

5. The proposed method can handle the group constraints between instances for practical applications.

- Target face association

1. We introduce a new video-based face identification task, where the probe video is only annotated once with a bounding box on the face of the target (subject of interest) in a video frame. The objective is to find the corresponding subject in the gallery based on this target-annotated probe video.

2. We propose a target face association method that establishes the association between face images of the target in multiple-shot videos. These associated face images are then utilized to create a robust representation of the target face.

- Face recognition using an outdoor camera network

  We describe a real-time video surveillance system consisting of four pan-tilt-zoom (PTZ) cameras at the University of Maryland campus. This system utilizes a master and slave camera framework to perform pedestrian tracking, face detection, and face recognition tasks.

1.6 Organization of the Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, we present a non-local cost aggregation method for character identification in TV-series. We discuss the problem of learning from ambiguously labeled data using matrix completion methods in Chapter 3. Then in Chapter 4, we present a solution to the problem of face association and video-based face identification. In Chapter 5, we summarize some recent face recognition techniques using outdoor camera networks

and present the design details of an outdoor camera network system at the University of Maryland campus. Finally, we conclude this dissertation and discuss future directions in Chapter 6.

Chapter 2: Character Identification in TV-series via Non-local Cost Aggregation

Character identification is an important task for preparing the metadata for a TV-series, since several applications, such as video summarization [3], analysis of character interactions [4], and shot retrieval [5, 6], require knowing the identities of humans in the scenes. Nevertheless, character identification in TV-series remains a challenging task since the video is usually unconstrained and the human pose varies.

Recently, non-local cost aggregation methods have been shown to yield good results in establishing dense stereo correspondence [7, 8]. This framework ensures that information is effectively utilized via non-local cost aggregation on a minimum spanning tree. Motivated by these works, we propose a non-local identification framework to recognize the identity of each face track and person track such that identities in the venue are consistently reported. Extending the non-local framework to solve the identification problem in a TV-series is not straightforward. Unlike pixels with identical modalities which line up in a planar graph structure, face tracks and person tracks own different modalities in the timeline. Besides, contextual information (e.g., unique presence of a subject) should be utilized to improve the identification performance.

We propose a unified approach for identifying the face and person tracks in a TV-series. We construct the track nodes to multiplex the modalities of face and clothing feature from face and person track, respectively. Our method possesses the capability to explore the video structure by constructing the minimum-distance spanning tree (MST) from the track nodes such that track nodes that are likely to have the same identity are adjacent to each other. A typical identification task assigns the identity such that the cost of each track node is minimized. By performing the non-local cost aggregation on MST, the identity assignment becomes more reliable via minimizing the aggregated cost, which allows the information from adjacent track nodes to be utilized as a whole. Furthermore, the unique presence of a subject in a venue is taken into account by minimizing the total aggregated cost of the track nodes with temporal dependency. Experimental results on TV-series datasets demonstrate the effectiveness of the proposed method.

The rest of the chapter is organized as follows. Section 2.1, we review some related work on face and person identification. In Section 2.2, we describe the proposed framework for character identification in TV-series. In Section 2.3, we demonstrate experimental results on two TV-series datasets. Section 2.4 concludes the chapter with a summary.

2.1 Related Work

Existing works use the names provided in the screenplay, speech identification [9], and attributes (e.g., gender) [10], to assist person identification. Furthermore,

analyzing the text information in subtitles has been used for person identification [11]. Nevertheless, several prior efforts [12,13] based on face clustering and tracking are not suitable for consistently identifying the characters in a TV-series due to shot variations and the occlusion of faces. Since the human body is more perceivable even when the face is occluded, person tracking [14] can provide additional advantages for person identification.

Our work is closely related to [9] that models the character appearances using Markov random field (MRF) representations of face and person tracks. Additional information, such as alternating shots and speaker identification, is utilized into their model to improve the identification performance. Their MRF framework relies on tracks with both face and clothing modalities, and transfers the face identification result to person tracks where faces cannot be authenticated due to occlusion or other reasons. However, this approach cannot guarantee that all the information in face and person tracks is utilized since the procedure of pre-clustering of clothing appearance and post-assignment of the identity to person tracks based on clustering results are performed separately.

2.2 The Proposed Framework

The block diagram of our approach is illustrated in Figure 2.1. First, the face and person tracks form track nodes if their bounding boxes co-occur with reasonable relative positions (Section 2.2.1). Track nodes with temporal dependency are then grouped into a knot (Section 2.2.2). We construct the MST from the k-knot graph

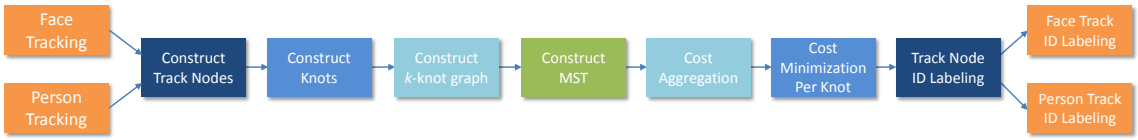


Figure 2.1: Block diagram of the character identification framework.

(Section 2.2.3) such that information can be conveyed across the track nodes in the video sequence. The cost of each track node is aggregated on the MST such that track nodes with similar appearance and temporal adjacency are more likely to have the same identity (Section 2.2.4). The identification problem is thus cast as a cost minimization problem per knot such that the uniqueness constraint is incorporated (Section 2.2.5). In the end, face and person tracks inherit the identity of the track node they are associated with.

2.2.1 Construction of Track Nodes

A track node can acquire feature modalities from both face and person tracks, which constitute a stronger representation than individual tracks. Besides, a track node typically has longer presence in timeline than its individual tracks, which allows the uniqueness constraint to be exploited over a longer period. Although a track node provides a unified representation, any erroneous matching of face and person track is irreversible. Besides, any error in the construction of track nodes causes immediate performance degradation in identification since face and person tracks inherit the identity of track node. Thus, we propose the following two-step procedure to construct the track nodes:

1) We use the Hungarian algorithm [15] to match the bounding boxes of face and person tracks in each frame. The Hungarian algorithm takes the cost matrix as input and outputs the matching status between the bounding boxes of face and person track in each frame. Each entry of the cost matrix corresponds to the distance between the center of face bounding box and the hypothesized center of the face according to the location of the person bounding box. If more than half the number of co-occurrence of bounding boxes of a face track and a person track are matched by the Hungarian algorithm, the face and person track are linked.

2) Each face and person track is initially treated as a track node. Two track nodes are merged into a larger track node if there is a link between two track nodes. This merging procedure is performed iteratively until it converges. Figure 2.3 illustrates a track node consisting of face and person tracks.

Track nodes can be categorized into three types: Face-body, face-only, and body-only track node. Face-body track nodes consist of both modalities from face and person tracks. Some track nodes only have a single modality, either from face or person track. Face-only track nodes appear when the human body cannot be detected. On the other hand, the body-only track nodes commonly appear when actors turn their bodies around. It is clear that face-body track nodes possess more information as compared to track nodes of a single modality.

2.2.2 Construction of Knots

Structural analysis of video can enhance the performance of identification. For instance, alternating shots [9] are common when filming conversations between two characters. This evidence can be utilized in identification by accumulating the decision for instances appearing in highly-correlated backgrounds. Nevertheless, the video structure usually depends on the media content, and it can be difficult to analyze a long shot. We propose to organize the track nodes into several knots. A knot is defined as the minimum set of track nodes with dependency in a temporal window such that there is no temporal dependency between any two knots. Track nodes can be organized into several knots using the following procedures:

1) We initialize a knot with a track node.

2) We iteratively augment other track nodes that share at least one common frame with any track node in this knot until it converges.

3) Construct another knot by going back to 1) until all the track nodes are organized.

In Figure 2.2(a), several track nodes are organized into three knots separated by dotted lines.

2.2.3 Construction of the k-Knot Graph

The k-nearest neighbor (k-NN) graph has been widely used to explore the latent structure of data. However, it does not consider the temporal structure of track nodes and the contextual information of the unique presence of a subject. Hence, we represent the structure of track nodes with an undirected k-knot graph to exploit

the contextual information of videos. The k-knot graph is constructed as follows: 1) Any two track nodes within a knot are connected with an edge if both track nodes do not share any common frame. This ensures two track nodes appearing in the same venue are set far apart while exploring the latent structure of track nodes. 2) As two track nodes from each of the two knots do not have temporal dependency, information can be transferred between them. Nevertheless, two track nodes become irrelevant if they are separated by a long temporal duration. In order to emphasize the information of a track node within a short temporal duration, a track node from the i^{th} knot will only be connected with track nodes from the $(i-k)^{th}$, $(i-k+1)^{th}$, ..., and $(i+k)^{th}$ knot. This allows the information to be successfully conveyed among track nodes for identification.

Figure 2.2(b) illustrates an example for constructing the k-knot graph from the track nodes. The distance between the i^{th} and j^{th} track node in the k-knot graph is defined as

$$d(i,j) = (1-\gamma)d_f(i,j) + \gamma d_p(i,j), \tag{2.1}$$

where γ controls the tradeoff between the distance induced by face and clothing modality. The $d_f(i,j)$ is the distance between the i^{th} and j^{th} track node induced by the face modality, which is defined as

$$d_f(i,j) = \begin{cases} \min\limits_{m \in F_i, n \in F_j} d_f(\mathbf{x}_i^m, \mathbf{x}_j^n) & \text{, if } F_i \neq \emptyset, F_j \neq \emptyset \\ d_f^{max} & \text{, otherwise,} \end{cases} \tag{2.2}$$

where F_i is the index set of face features in the i^{th} track node, and $\mathbf{x}_i^m$ represents

the m^{th} face feature vectors of the i^{th} track node. If the track node i or j lack the modality from face tracks, $d_f(i,j)$ will be set equal to d_f^{max}, which is the maximum value of $d_f(i,j)$. $d_f(\mathbf{x}_1, \mathbf{x}_2)$ denotes the cosine similarity [16] between $\mathbf{x}_1$ and $\mathbf{x}_2$, and sophisticated metrics, such as the one discussed in [17], can be utilized to improve the performance.

The clothing feature is the RGB color histogram computed from the bounding box corresponding to the torso in the person track. Similarly, we define the distance between the i^{th} and j^{th} track node induced by the clothing modality as

$$d_p(i,j) = \begin{cases} \min\limits_{m \in P_i, n \in P_j} d_p(\mathbf{h}_i^m, \mathbf{h}_j^n) & \text{, if } P_i \neq \emptyset, P_j \neq \emptyset, \\ d_p^{max} & \text{, otherwise,} \end{cases} \tag{2.3}$$

where P_i is the index set of clothing features in the i^{th} track node. The $\mathbf{h}_i^m$ denotes the m^{th} clothing feature vector of the i^{th} track node. $d_p(\mathbf{h}_1, \mathbf{h}_2)$ represents the chi-squared distance between histogram feature $\mathbf{h}_1$ and $\mathbf{h}_2$. Note that $d_p(i,j)$ will be set equal to the maximum distance d_p^{max} if track node i or j lack the clothing modality.

2.2.4 Construction of the Minimum-distance Spanning Tree (MST)

The construction of MST automatically removes the unwanted edges of large distance such that the total distance of the spanning tree is minimized. Motivated by this fact, we construct the MST from the k-knot graph to explore the structure of track nodes. In Figure 2.2(c), we observe that track nodes of the same identity are closer in MST since edges of large distance are removed during the construction of MST. Note that the distance between two track nodes in MST is defined as the summation of distances along the edges connecting these two track nodes. Hence,

we define $D(i,j)$ as the distance between the i^{th} and j^{th} track node in MST. Note that $D(i,j) = d(i,j)$ if the i^{th} and j^{th} track node are directly connected by an edge in MST. We define the similarity between the i^{th} and j^{th} track node as

$$S(i,j) = \exp\left(-\frac{D(i,j)}{\sigma}\right), \tag{2.4}$$

where σ is the parameter to adjust the similarity.

Let $C_i(y)$ represent the cost for the i^{th} track node if it is treated as identity $y \in \mathcal{C}$, where $\mathcal{C} = \{1, 2, \ldots, c\}$ is the identity set. The modeling of $C_i(y)$ will be discussed in Section 2.2.6. Following the non-local cost aggregation framework presented in [7], the aggregated cost of the i^{th} track node is computed by

$$C_i^A(y) = \sum_j S(i,j) C_j(y) = \sum_j \exp\left(-\frac{D(i,j)}{\sigma}\right) C_j(y). \tag{2.5}$$

The aggregation procedure can be treated as a filtering operation, where each track node contributes to the task of identification via similarity weighting. Identification becomes robust since information from adjacent track nodes is utilized as a whole for determining the identity. Although the cost aggregation in (2.5) requires the weighted summation across all the track nodes, Yang [7] provides a linear time exact algorithm to significantly reduce the computational burden.

2.2.5 Cost Minimization per Knot

The identity of the i^{th} track node can be obtained by assigning the identity that minimizes its aggregated cost in (2.5). Since a knot consists of several track nodes with temporal dependency, the identity of track nodes from a knot should be

jointly determined. The solution can be obtained by enumerating all combinations of labeling that are consistent with the uniqueness constraint such that the aggregated cost of knot is minimized. Hence, we can predict the identities of track nodes in the j^{th} knot by solving

$$\hat{\mathbf{y}}_j = \arg\min_{\mathbf{y}\in\mathcal{Y}_j} \sum_{i\in O_j} C_i^A(y_i), \tag{2.6}$$

where O_j is the set containing the indices of track nodes in the j^{th} knot. The identities of track nodes in the j^{th} knot form a column vector $\mathbf{y}$, and $\mathcal{Y}_j$ is the set consisting of all the combinations of identities that satisfy the unique presence of an identity. This combinatorial problem can be solved by an optimization procedure based on relaxation technique discussed in [9]. Once the identities of track nodes are determined, the face and person tracks inherit the identity of the track node it is associated with.

2.2.6 Modeling the Cost Function

The cost function returns the amount of the deviation from the designated subject. Herein, we define the cost for treating the i^{th} track node as identity y as

$$C_i(y) = \begin{cases} -r_i(y), & \text{if} \quad F_i \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Note that the i^{th} track node has cost equal to 0 if it lacks the face modality, i.e. $F_i = \emptyset$. Hence, track nodes that miss face modality will passively receive the information propagated from adjacent track nodes. Without loss of generality, the

unknown class is regarded as the c^{th} class, and $r_i(y)$ in (2.7) is modeled as

$$r_i(y) = \begin{cases} \frac{1}{|F_i|} \sum_{m \in F_i} \Phi_y(\mathbf{x}_i^m), y \in \{1, 2, \ldots, c-1\}, \\ \frac{1}{|F_i|} \sum_{m \in F_i} \lambda \min_{j \neq c} \left(1 - \Phi_j(\mathbf{x}_i^m)\right), y = c, \end{cases} \tag{2.8}$$

where $\Phi_y(\mathbf{x})$ returns the probabilistic output from a support vector machine (SVM) [18, 19]. We follow the setting in [9] to train the SVMs with a second-order polynomial kernel. The training data of the first $(c-1)^{th}$ classes are used to train $c-1$ classifiers using one-versus-all SVM. Note that we do not explicitly model the unknown class since the number of face tracks corresponding to the unknown class is usually insufficient to model the unknown class. Hence, we use the minimum complementary of the probabilistic output among $c-1$ subjects to model the likelihood of the unknown class. However, this excessively biases towards the unknown class as the minimum complementary of the probabilistic output can be large for unseen data. We use λ to adjust the likelihood of the unknown class, and λ is obtained from classifying the validation data such that the classification accuracy is maximized. The validation data consists of a subset of training samples from major characters and all the training samples of the unknown class.

2.3 Experimental Results

In this section, we perform experiments on the TV-series datasets in Section 2.3.1 and discuss the results in Section 2.3.2.

2.3.1 Evaluation on TV-series Datasets

We use two datasets provided by the authors of [9,20] for evaluation.[1] The first dataset consists of 6 episodes of *Big Bang Theory* (BBT), and the second dataset consists of 6 episodes of *Buffy the Vampire Slayer* (BF). We use the face features readily provided in these datasets. The dimension of the feature vector is 240, and the feature coefficients are computed using the discrete cosine transform (DCT) from face regions of 48×64 pixels. The BBT dataset provides face and person tracks, while the BF dataset only provides face tracks. Moreover, only 22 % (recall) face tracks are labeled via matching the name of the transcript with the face track that is speaking [20, 21]. These face tracks are weakly labeled with 87 % accuracy (precision) due to the falsely detected speaking face and the mismatch of transcripts. Note that we do not specifically handle the potentially erroneous labeling situation and use all the available labels for training. More sophisticated methods, reported in [20, 22], can be utilized to further improve the identification performance.

The identification accuracy of face (person) track is computed as the number of correctly identified face (person) tracks over the total number of face (person) tracks in each episode. For comparison, we follow the same setting reported in [20]. There are 11 and 28 subjects in the BBT dataset and BF dataset, respectively. Each dataset has an additional unknown class. Characters that do not belong to any subjects are regarded as belonging to the unknown class, and the uniqueness constraint is not applied to the unknown class. Since the BF dataset does not

[1] Dataset is available at http://cvhci.anthropomatik.kit.edu/projects/mma.

provide the person track, we compute the clothing features from a hypothesized rectangular region below the face. A similar procedure is also reported in [23] to extract the clothing features. Hence, the BF dataset demonstrates another scenario where person tracks are not available, and each face track is trivially treated as a track node. Throughout all the experiments, we use $\gamma = 0.8$, $\sigma = 0.1$, and $k = 10$.

2.3.2 Discussion on the Experiments

We compare our method with the person identification framework based on MRF [9], which takes the probabilistic output from a trained classifier using semi-supervised learning with constraints (SSLC) [20]. Our trained classifier is denoted as SVM. The performance of identification evaluated on "track node" and "track node with cost aggregation" is denoted as TN and TN+CA, respectively. Considering the uniqueness constraint, we denote TN+CA+K as "TN+CA with cost minimization per knot". Based on the experimental results, we make the following observations:

1. In Table 2.1, our SVM gives identification accuracy of 78.21% for the BBT dataset. Specifically, the construction of track nodes provides 0.85% improvement. It provides a slight improvement since face tracks belonging to the same track node can be fused for reliable identification. The cost aggregation procedure provides additional 4.49% improvement since the filtering operation of cost aggregation can propagate the information to track nodes without face modality and suppress the impact of noisy instances. Considering the uniqueness constraint, we use (2.6) to jointly determine the identities of track nodes of each knot. Overall, our method

(TN+CA+K) outperforms the MRF framework with SSLC (SSLC+MRF) by 3.48%. The confusion matrix for the identification of face tracks in the BBT dataset is presented in Figure 2.4, which shows that our method is effective in classifying all the characters including the guest characters. Note that Doug and Summer are not correctly identified since no weakly-labeled face track is associated with these two subjects for training.

2. In Table 2.2, TN + CA significantly improves the identification accuracy of person tracks over TN, which shows that the identification result of face tracks is successfully transferred to the person track via the cost aggregation on MST. When the uniqueness constraint is considered, TN+CA+K attains the identification accuracy of 86.66%.

3. Since the BF dataset does not provide person tracks, each face track is treated as a track node. Therefore, the performance of SVM and TN are identical. In Table 2.1, our method (TN + CA) achieves the identification accuracy of 69.39%. However, enforcing the uniqueness constraint gives only a minor improvement. One of the reasons is that the video structure of BF usually has one or two characters in the scene, which does not provide as much contextual information as in BBT. Our method (TN + CA + K) outperforms SSLC and SVM by 3.71% and 4.26 %, respectively.

4. The performance of person identification task depends on the recognizing accuracy of the face classifier. In order to fairly compare with [9], we evaluate the identification accuracy of person tracks by providing the groundtruth for face tracks. Following the protocol in [9], we use the BBT dataset to evaluate the character iden-

tification component. Only the five main characters and the additional unknown class are evaluated. As we can observe in Table 2.3, our method (TN+CA+K) achieves 4.5% improvement over [9]. This shows that our method performs better than [9] since the information from face and person tracks is utilized as a whole for identification.

5. Table 2.4 shows the statistics of track nodes in the BBT dataset. In order to investigate the quality of track nodes, we verify whether the face and person tracks in a track node have the same identity using the groundtruth. A track node with any inconsistent identities among its face and person tracks is regarded as an erroneous track node. It is clear that erroneous track nodes only account for a small portion (0.4%) of all the track nodes. In contrast to the MRF framework proposed in [9] where the identities of face tracks are first recognized and transferred to person tracks without visible face based on the affinity of the clothing appearance, our method ensures that all the information is utilized as a whole. Although body-only track nodes lack face modality, they serve as the relay for propagating the inference of other track nodes. Moreover, the duration of the track node accounts for the temporal appearance of an identity in the timeline, and thus the pairwise constraint between track nodes is generally stronger than just the face or person track alone.

2.4 Summary

We proposed a unified framework for character identification in a TV-series. We constructed the track nodes from face and person tracks, and used the track

Table 2.1: Identification accuracy of face tracks in BBT and BF datasets.

Episode	BBT-1	BBT-2	BBT-3	BBT-4	BBT-5	BBT-6	BBT-Avg.	BF-1	BF-2	BF-3	BF-4	BF-5	BF-6	BF-Avg.
SSLC [20]	89.23	89.20	78.47	76.59	75.09	68.05	79.44	71.99	61.27	66.60	67.07	69.59	61.72	66.37
SSLC + MRF [9, 20]	95.18	94.16	77.81	79.35	79.93	75.85	83.71	–	–	–	–	–	–	–
SVM	87.94	85.84	77.81	76.25	72.76	68.66	78.21	69.63	62.20	64.20	67.07	69.34	62.45	65.82
TN	90.35	87.26	78.47	77.11	72.04	69.15	79.06	69.63	62.20	64.20	67.07	69.34	62.45	65.82
TN + CA	92.28	91.15	82.22	84.85	78.85	71.95	83.55	74.08	62.31	67.81	72.10	75.95	64.11	69.39
TN + CA + K	94.21	92.39	84.01	87.78	83.15	81.59	87.19	75.65	64.38	66.70	72.81	76.97	63.93	70.08

Table 2.2: Identification accuracy of person tracks in the BBT dataset.

Episode	1	2	3	4	5	6	Avg.
TN	78.54	73.04	75.58	63.77	64.07	61.49	69.42
TN + CA	89.12	87.77	83.67	80.97	78.31	71.01	81.81
TN + CA + K	91.80	89.81	87.71	86.61	82.95	81.09	86.66

Table 2.3: Identification accuracy of person tracks given the groundtruth identities of face tracks in the BBT dataset.

Episode	1	2	3	4	5	6	Avg.
MRF [9]	98.3	89.9	94.8	89.1	85.3	88.5	91.0
TN	86.0	82.6	86.9	78.2	79.1	76.7	81.6
TN + CA	95.7	93.4	96.4	91.5	88.9	85.2	91.8
TN + CA + K	97.8	96.9	97.4	94.2	94.5	92.5	95.5

Table 2.4: Statistics of the tracks, track nodes, and knots in Episode 1-6 of BBT.

Episode	1	2	3	4	5	6
# Face tracks	622	565	613	581	558	820
# Person tracks	671	638	643	657	604	883
Our track nodes (TNs)						
# Face-body TNs	527	469	476	481	424	604
# Face-only TNs	14	21	66	43	80	156
# Body-only TNs	142	168	158	174	174	262
# All TNs	683	658	700	698	678	1022
# Erroneous TNs	0	1	1	2	6	11
# Knots	362	348	375	330	281	316

nodes to serve as the basic unit in constructing the MST. Hence, track nodes with similar appearance are adjacent in MST. Then non-local cost aggregation was performed on the MST, which serves as a filtering operation to suppress the impact of noisy instances and provides the inference to track node without face modality. Considering the unique presence of a subject, the identities of track nodes with temporal dependency was jointly determined by minimizing the aggregated cost of those track nodes. Experimental results confirm the effectiveness of our method.

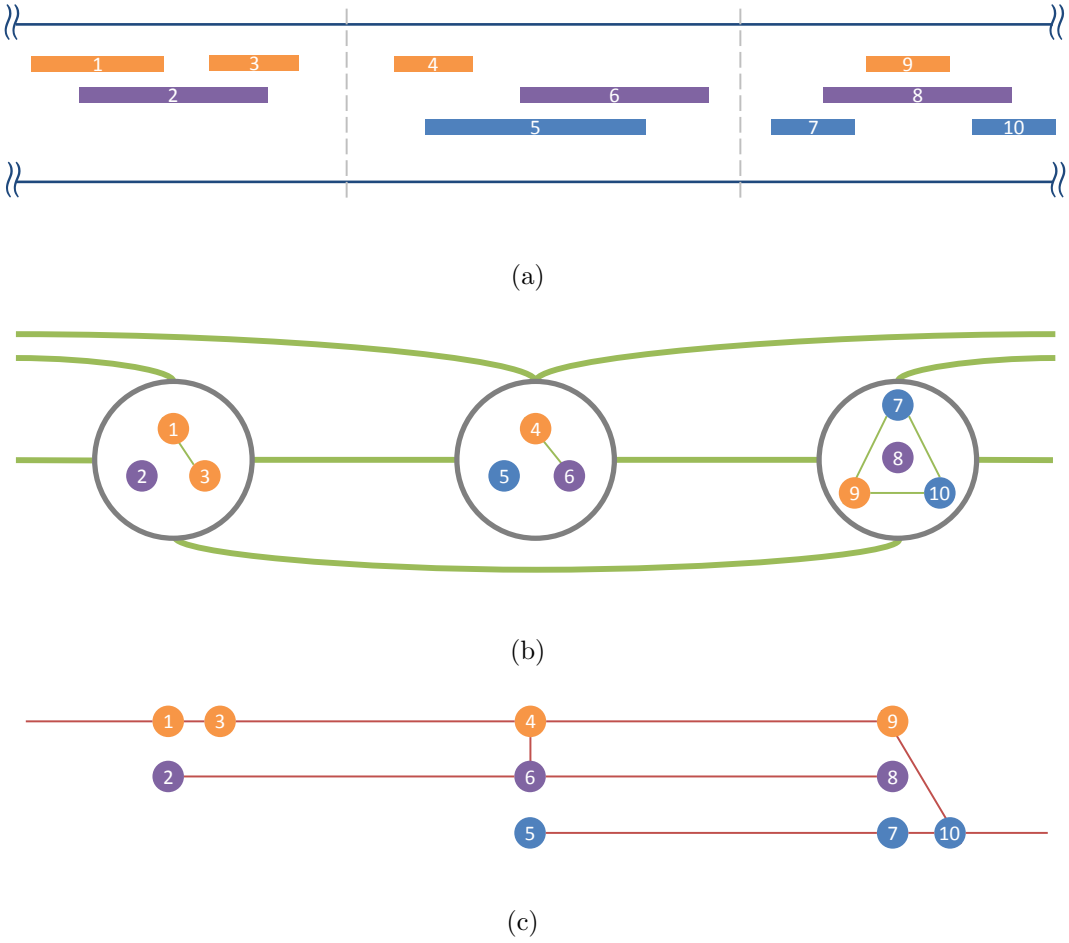


Figure 2.2: Construct the MST from track nodes of three identities (color-encoded as orange, purple, and blue). (a) Knot construction: Track nodes are organized into three knots (separated by the dotted lines). (b) k-knot graph: The thin green lines represent the edges between the track nodes in a knot, and any pair of track nodes from each of the two knots linked by the bold green lines is connected by an edge. (c) MST: Edges of large distances in the k-knot graph are removed. Hence, track nodes of the same identity are more likely to be connected since their associated edges have relatively small distances.

Figure 2.3: Face tracks (blue and green) and a person track (red) are merged into one track node.

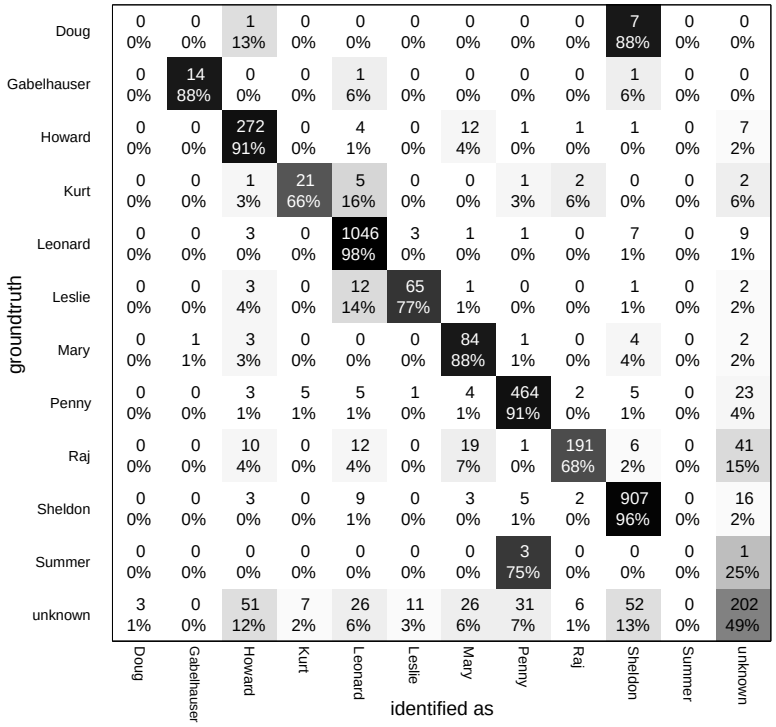


Figure 2.4: Confusion matrix over Episode 1-6 of BBT for TN + CA + K.

Chapter 3: Learning from Ambiguously Labeled Face Images

Learning a classifier for naming a face requires a large amount of labeled face images and videos. However, labeling face images is expensive and time-consuming due to significant amount of human efforts involved. As a result, brief descriptions such as tags, captions and screenplays accompanying the images and videos become important for training the classifiers. Although such information is publicly available, it is not as explicitly labeled as human annotations. For instance, names in the caption of a news photo provide possible candidates for faces appearing in the image [24, 25] (see Figure 3.1). The names in the screenplays are only weakly associated with faces in the shots [21]. The problem in which instead of a single label per instance, one is given a candidate set of labels, of which only one is correct is known as ambiguously labeled learning[1] [10, 26–29].

In recent years, the problem of completing a low-rank matrix with missing entries has gained a lot of attention. In particular, matrix completion methods have been shown to produce good results for multi-label image classification problems [30], [31]. In these methods, the underlying assumption is that the concatenation of feature vectors and their labels produce a low-rank matrix. Our work is motivated by

[1] also known as partially labeled learning and superset label learning

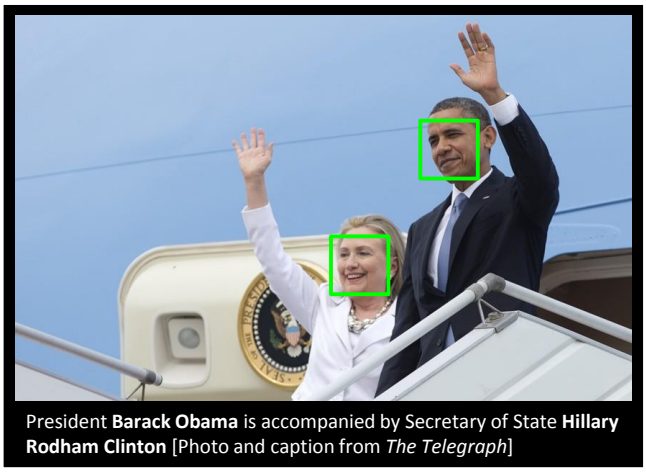


Figure 3.1: The names in the captions are not explicitly associated with the face images appeared in the news photo.

these works. The proposed method, Matrix Completion for Ambiguity Resolution (MCar), takes the heterogeneous feature matrix, which is the concatenation of the labeling matrix and feature matrix, as input. We first show that the heterogeneous feature matrix is ideally low-rank in the absence of noise. This in turn, allows us to convert the labeling problem as a matrix completion problem by pursuing the underlying low-rank matrix of the heterogeneous feature matrix. In contrast to multi-label learning, ambiguous labeling provides the clue that one of the labels in the candidate label set is the true label. This knowledge is utilized to regularize the labeling matrix in the heterogeneous feature matrix. This is essentially the main difference between our work and some of the previously proposed matrix completion techniques [30], [31].

Although ambiguous learning techniques can take advantage of large-scale and diverse ambiguously labeled data, most of the methods cannot properly handle the

labeling imbalance that is often present in publicly available training data. For instances, celebrities and leading actors usually dominate (appear more frequently) in the candidate label sets, and these majority labels can easily bias the results of ambiguity resolving methods. As the proposed method relies on low-rank approximation of the heterogeneous feature matrix, heterogeneous feature vectors associated with those majority labels can dominate the process of low-rank approximation and thus bias the recovery of the labeling matrix. We propose the weighted MCar (WMCar) to overcome the labeling imbalance in ambiguously labeled data. Unlike conventional instance weighting techniques [32] that assign unequal instance weight to the cost function of instances, WMCar performs unequal column-wise weighting on the heterogeneous feature vectors. Therefore, a heterogeneous feature vector associated with majority labels will contribute less to the process of low-rank approximation than that associated with minority labels.

The column-wise weighting in WMCar can be computed by estimating the groundtruth label distribution from the recovered labeling matrix, but the recovered labeling matrix is not accessible without applying WMCar to resolve the ambiguity in the original labeling matrix. Nevertheless, iteratively updating the column-wise weighting and recovering the labeling matrix with WMCar is not reliable (see iterative WMCar in Figure 3.15). An explanation is that there is some unresolved ambiguity in the soft labeling matrix recovered by WMCar. The remaining ambiguity (noise) can be detrimental to the iterative process as we iteratively update WMCar by substituting the labeling matrix with the recovered one from the previous iteration. Hence, we propose the Iterative Candidate Elimination (ICE) procedure

to iteratively eliminate the least likely candidates from a portion of the ambiguously labeled data. This procedure iteratively suppresses the noise in the recovered labeling matrix and thus yields a better performance in the next iteration of WMCar. Although WMCar with ICE is an iterative approach, it is fundamentally different from previously suggested iterative methods [28, 33, 34]. Unlike previous works that iteratively construct class-specific models and update the labels, the iterative process of ICE is effective in sequential noise suppression. Besides, WMCar concatenates the labels and features as a heterogeneous matrix to recover the labels in each iteration. This ensures that the information in the ambiguously labeled data is used as a whole in recovering the true labels.

Moreover, we generalize MCar to include the labeling constraints between the instances for practical applications. For instances, two persons in a news photo should not be identified as the same subject even though both of them are ambiguously labeled in the caption. As shown by the recent success in low-rank matrix recovery [35], several prior works have developed robust methods for classification [36], [37]. The proposed method inherits the benefit of low-rank matrix recovery and possesses the capability to resolve the label ambiguity via low-rank approximation of the heterogeneous matrix. As a result, our method is more robust compared to some of the existing discriminative ambiguous learning methods [10, 38]. The disambiguated labels from MCar are then used to learn a supervised learning classifier, which can be used to classify new data.

We use the following notations in this chapter. The matrix element $a_{i,j}$ denotes the entity in the i^{th} row and j^{th} column of matrix $\mathbf{A}$. $\mathbf{1}_n$ represents a column vector

of size $n \times 1$ consisting of 1's as its entries. $\|\cdot\|_1$ and $\|\cdot\|_0$ denote the ℓ_1 norm and ℓ_0 norm, respectively. The Frobenius norm and the nuclear norm of $\mathbf{A}$ are defined as $\|\mathbf{A}\|_F = \left(\sum_{i,j}(a_{i,j})^2\right)^{\frac{1}{2}}$ and $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A})$, respectively where σ_i is the i^{th} singular value of $\mathbf{A}$. $(\cdot)^T$ denotes transposition operation. $|S|$ returns the cardinality in set S. $\mathcal{S}_a[b] = \text{sgn}(b)\max(|b| - a, 0)$ is the shrinkage operator. The concatenation of matrix $\mathbf{A}$ and $\mathbf{B}$ is defined as $\begin{bmatrix}\mathbf{A}\\ \\ \mathbf{B}\end{bmatrix} = [\mathbf{A}; \mathbf{B}]$.

The rest of this chapter is organized as follows. In Section 3.1, we review some related work on ambiguously labeled learning methods. Section 3.2 describes the proposed MCar and WMCar. The optimization procedure for WMCar is described in Section 3.3. Section 3.4 describes the ICE procedure in detail. Section 3.5 presents the extension of MCar for incorporating the constraint between instances. In Section 3.6, we demonstrate the results on synthesized as well as real-world ambiguously labeled datasets. Finally, Section 3.7 concludes this work with a brief summary and discussion.

3.1 Related Work

Various methods have been proposed in the literature for dealing with ambiguously labeled data. Some of these methods propose Expectation Maximization (EM)-like approaches to alternately disambiguate the labels and learn a discriminative classifier [39, 40]. Non-parametric methods have also been used to resolve the ambiguity by leveraging the inductive bias of learning methods [26]. For the ambiguously labeled training data the actual loss of mislabeling is not explicit. As a result,

it is difficult to learn an effective discriminative model. Cour *et al.* [10,41] proposed the partial 0/1 loss function for ambiguous labeling, which is a tighter upper bound for the actual loss as compared to the 0/1 loss [42]. Subsequently, a discriminative classifier can be learned from the ambiguous labels by minimizing the partial 0/1 loss. Several works have improved the learning of partial labels with the modeling of partial loss [43], error-correcting output codes [44], and iterative label propagation [45]. Liu *et al.* [27] proposed to learn a conditional multinomial mixture model for predicting the actual label from ambiguous labels. Several dictionary-based methods have also been proposed for handing partially labeled datasets [28,34,46]. In particular, an EM-based dictionary learning approach was proposed in [28], where a confidence matrix and dictionary are updated in alternating iterations. Although dictionary-based methods are robust to occlusions and noise, the EM-based approach can be very sensitive to the selection of initial dictionary and also may suffer from suboptimal performance.

Luo *et al.* [38] generalize the ambiguously labeled learning problem addressed in [10] from single instances to a group of instances. The ambiguous loss considers the association between the group of identities and the candidate label vectors. The pairwise constraint between the instances (e.g. unique appearance of a subject) is accounted for when generating the candidate label vectors. Furthermore, Zeng *et al.* [33] use a Partial Permutation Matrix (PPM) to associate the identities in a group with ambiguous labels. The pairwise constraint is encoded by restricting the structure of PPM. Assuming that instances of the same subject inferred by PPM can ideally form a low-rank matrix, the actual identity of an instance can be predicted

by alternatively updating the low-rank subspace and PPM. Xiao *et al.* [47] associate the identities in a group from ambiguous labels by minimizing the summation of the discriminative affinities in a group, where the affinities are learned from the low-rank reconstruction coefficient matrix and the weak supervision of ambiguous labels.

Recently, learning from weak annotations of labeling imbalance has received significant attention [48, 49]. Chen *et al.* [50] have employed the part-versus-part decomposition [51] to overcome the data imbalance in multi-label learning. Charte *et al.* [52] propose several methods to resample the multi-label training data to compensate the imbalance level. Wu *et al.* [53] incorporate the class cardinality bound constraints to deal with class imbalance. Although several prior works have addressed the issue of imbalanced data in the context of multi-label learning, the labeling imbalance in ambiguously labeled data remains to be investigated. We propose to estimate the groundtruth label distribution from ambiguous labels. With the estimated groundtruth label distribution, the instance weight of WMCar can be computed to deal with labeling imbalance.

3.2 The Proposed Framework

The ambiguously labeled data is denoted as $\mathcal{L} = \{(\mathbf{x}_j, L_j),\ j = 1, 2, \ldots, N\}$, where N is the number of instances. There are c classes, and the class labels are denoted as $\mathcal{Y} = \{1, 2, \ldots, c\}$. Note that $\mathbf{x}_j$ is the feature vector of the j^{th} instance, and its candidate labeling set $L_j \subseteq \mathcal{Y}$ consists of candidate labels associated with the j^{th} instance. The true label of the j^{th} instance is $l_j \in L_j$. In other words, one

of the labels in L_j is the true label of $\mathbf{x}_j$. The objective is to resolve the ambiguity in $\mathcal{L}$ such that each predicted label $\hat{l}_j$ of $\mathbf{x}_j$ matches its true label l_j. We associate the candidate labeling set L_j with a soft labeling vector $\mathbf{p}_j$, where $p_{i,j}$ indicates the probability that instance j belongs to class i. This allows us to quantitatively assign the likelihood of each class the instance belongs to if such information is provided. Given the ambiguous label of the j^{th} instance, we assign each entry of $\mathbf{p}_j$ as

$$\begin{cases} p_{i,j} = (0,1] & \text{if } i \in L_j, \\ p_{i,j} = 0 & \text{if } i \notin L_j, \end{cases} \quad j = 1, 2, \ldots, N, \tag{3.1}$$

where $\sum_{i=1}^{c} p_{i,j} = 1$. Without any prior knowledge, we assume equal probability for each candidate label. Let $\mathbf{P} \in \mathbb{R}^{c \times N}$ denote the ambiguous labeling matrix with $\mathbf{p}_j$ in its j^{th} column. With this, one can model the ambiguous labeling as

$$\mathbf{P}^0 = \mathbf{P} - \mathbf{E}_P, \tag{3.2}$$

where $\mathbf{P}^0$ and $\mathbf{E}_P$ denote the true labeling matrix and the labeling noise, respectively. The j^{th} column vector of $\mathbf{P}^0$ is $\mathbf{p}_j^0 = \mathbf{v}_{l_j}$, where $\mathbf{v}_{l_j}$ is the canonical vector corresponding to the 1-of-K coding of its true label l_j.

Similarly, assuming that the feature vectors are corrupted by some noise or occlusion, the feature matrix $\mathbf{X}$ with $\mathbf{x}_j$ in its j^{th} column can be modeled as

$$\mathbf{X}^0 = \mathbf{X} - \mathbf{E}_X, \tag{3.3}$$

where $\mathbf{X} \in \mathbb{R}^{m \times N}$ consists of N feature vectors of dimension m, $\mathbf{X}^0$ represents the feature matrix in the absence of noise and $\mathbf{E}_X$ accounts for the noise. Concatenating (3.2) and (3.3), we obtain a unified model of ambiguous labels and feature vectors,

which can be expressed as

$$\begin{bmatrix} \mathbf{P}^0 \\ \mathbf{X}^0 \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix}. \tag{3.4}$$

Let

$$\mathbf{H}_{obs} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} \text{ and } \mathbf{E} = \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix} \tag{3.5}$$

denote the heterogeneous feature matrix and its noise, respectively. If we can show that $\mathbf{H}_{obs}$ is a low-rank matrix in the absence of noise, then we can use matrix completion methods for resolving the ambiguity in labeling. In the following section, we investigate the low-rank property of $\mathbf{H}_{obs}$.

3.2.1 Exploiting the Rank of $\mathbf{H}_{obs}$

The column vectors of $\mathbf{X}_0$ can be partitioned into sets $S_1, S_2, \ldots, S_c$ based on their true labels. We assume that the elements of S_k form a convex hull C_k of n_k vertices. It is clear that $n_k \leq |S_k|$. The representative matrix of the k^{th}class, $\mathbf{D}_k \in \mathbb{R}^{m \times n_k}$, consists of vertices of C_k as its column vectors, and each column vector is treated as a representative of the k^{th}class. Therefore, according to the definition of a convex hull, a noise-free instance $\mathbf{x}_j^0$ from class k ($\mathbf{x}_j^0 \in C_k$) can be represented as

$$\mathbf{x}_j^0 = \mathbf{D}_k \mathbf{a}_{k,j}, \text{ where } \mathbf{a}_{k,j}^T \mathbf{1}_{n_k} = 1, \mathbf{a}_{k,j} \in \mathbb{R}_+^{n_k \times 1}. \tag{3.6}$$

Note that $\mathbf{a}_{k,j} \in \mathbb{R}_+^{n_k \times 1}$ is the coefficient vector associated with the representative matrix of the k^{th} class. As the true label of an instance is not known in advance,

we can represent $\mathbf{x}_j^0$ as

$$\begin{aligned} \mathbf{x}_j^0 &= \mathbf{D}\mathbf{q}_j, \\ \mathbf{D} &= [\mathbf{D}_1 \ \mathbf{D}_2 \ \cdots \ \mathbf{D}_c], \\ \mathbf{q}_j &= [\mathbf{a}_{1,j}^T \ \mathbf{a}_{2,j}^T \ \cdots \ \mathbf{a}_{c,j}^T]^T, \ \mathbf{q}_j^T\mathbf{1} = 1, \end{aligned} \tag{3.7}$$

where $\mathbf{D} \in \mathbb{R}^{m\times(\sum_{i=1}^c n_i)}$ is the collective representative matrix, and $\mathbf{q}_j \in \mathbb{R}_+^{(\sum_{i=1}^c n_i)\times 1}$ is the associated coefficient vector.

According to (3.7), we can decompose $\mathbf{X}^0$ as

$$\mathbf{X}^0 = \mathbf{D}\mathbf{Q}. \tag{3.8}$$

The coefficient matrix $\mathbf{Q}$ in (3.8) is not unique as column vectors of $\mathbf{D}$ are not necessarily linearly independent. However, we assume that an ideal decomposition $\mathbf{X}^0 = \mathbf{D}\mathbf{Q}^*$ satisfies the following condition

$$\begin{aligned} \mathbf{x}_j^0 = \mathbf{D}\mathbf{q}_j^*, \ \text{where } \mathbf{a}_{k,j}^{*T}\mathbf{1}_{n_k} = 1, \ \mathbf{x}_j^0 \in S_k, \\ \mathbf{a}_{l,j}^{*T}\mathbf{1}_{n_l} = 0, \ l \neq k, \end{aligned} \tag{3.9}$$

which implies that $\mathbf{x}_j^0$ is exclusively represented by $\mathbf{D}_k$ even though it is possible that it can be written as a linear combination of any other vertices from different classes.

With this, we can recover the true labels from

$$\mathbf{P}^0 = \mathbf{T}\mathbf{Q}^*, \tag{3.10}$$

where $\mathbf{T} = [\mathbf{v}_1\mathbf{1}_{n_1}^T \ \mathbf{v}_2\mathbf{1}_{n_2}^T \ \cdots \ \mathbf{v}_c\mathbf{1}_{n_c}^T]$ accumulates the coefficients associated with each matrix representative. Hence, the coefficient vector of dimension $\sum_{i=1}^c n_i$ is

converted into labeling vector of dimension c. Concatenating $\mathbf{P}^0 = \mathbf{T}\mathbf{Q}^*$ and $\mathbf{X}^0 = \mathbf{D}\mathbf{Q}^*$, we further represent (3.4) as

$$\begin{bmatrix} \mathbf{P}^0 \\ \mathbf{X}^0 \end{bmatrix} = \begin{bmatrix} \mathbf{T} \\ \mathbf{D} \end{bmatrix} \mathbf{Q}^*. \tag{3.11}$$

It is clear that

$$\begin{aligned} \operatorname{rank}(\begin{bmatrix} \mathbf{P}^0; \mathbf{X}^0 \end{bmatrix}) &\leq \min\left(\operatorname{rank}(\begin{bmatrix} \mathbf{T}; \mathbf{D} \end{bmatrix}), \operatorname{rank}(\mathbf{Q}^*)\right) \\ &\leq \min\left(c+m, \sum_{k=1}^{c} n_k, N\right). \end{aligned} \tag{3.12}$$

Since the representatives in $\mathbf{D}$ only account for a subset of data samples, it is clear that $\sum_{k=1}^{c} n_k \leq N$. Therefore,

$$\operatorname{rank}(\begin{bmatrix} \mathbf{P}^0; \mathbf{X}^0 \end{bmatrix}) \leq \min\left(c+m, \sum_{k=1}^{c} n_k\right). \tag{3.13}$$

The rank of $[\mathbf{P}^0; \mathbf{X}^0]$ is at most $\sum_{k=1}^{c} n_k$ if the dimension of feature vectors m is not less than the number of representatives in $\mathbf{D}$, i.e. $\sum_{k=1}^{c} n_k \leq m$. Hence, $[\mathbf{P}^0; \mathbf{X}^0]$ has a relatively smaller rank than N in the case of $N >> \min\left(c+m, \sum_{k=1}^{c} n_k\right)$. From the above rank analysis and (3.4), we arrive at the following proposition:

Proposition 1. *The heterogeneous feature matrix $\mathbf{H}_{obs}$ is low-rank in the absence of noise.*

Note that a similar result is also reported in [54] without making the convex hull assumption.

3.2.2 Matrix Completion for Ambiguity Resolution

According to (3.10), the true labeling matrix $\mathbf{P}^0$ can be recovered if $\mathbf{D}$ and $\mathbf{Q}^*$ are available. Nevertheless, obtaining $\mathbf{D}$ and $\mathbf{Q}^*$ based on the observed $\mathbf{P}$ and $\mathbf{X}$ is

intractable by solving a matrix decomposition problem

$$\min_{\mathbf{T},\mathbf{D},\mathbf{Q}} \left\| \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} - \begin{bmatrix} \mathbf{T} \\ \mathbf{D} \end{bmatrix} \mathbf{Q} \right\|_F^2, \tag{3.14}$$

subject to the conditions specified in (3.9)-(3.11). Following [30], we propose to resolve the ambiguity by recovering the underlying low-rank structure of the heterogeneous feature matrix. Hence, we transform the matrix decomposition problem to a matrix completion problem. For the ease of presentation, we start with solving a label assignment problem assuming that $\mathbf{X}$ is noise-free, i.e. $\mathbf{X} = \mathbf{X}^0$. The predicted labeling matrix $\mathbf{Y}$ can be estimated by solving the following rank minimization problem

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{E}_P} \ & \operatorname{rank}\left(\begin{bmatrix} \mathbf{Y} \\ \mathbf{X}^0 \end{bmatrix} \right) \\ \text{s.t.} \ & \begin{bmatrix} \mathbf{Y} \\ \mathbf{X}^0 \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X}^0 \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P \\ \mathbf{0} \end{bmatrix}, \\ & \mathbf{y}_j \in \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c\}, j = 1, 2, \ldots, N, \\ & y_{i,j} = 0 \text{ if } i \notin L_j \ \forall j. \end{aligned} \tag{3.15}$$

The problem is to complete the labeling matrix $\mathbf{Y}$ via pursuing a low-rank matrix $[\mathbf{Y}; \mathbf{X}^0]$ subject to the constraints given by the ambiguous labels. The first constraint defines the feasible region of label assignment and the second constraint implies that an instance can only be labeled among its candidate labels. We cannot guarantee that the optimal solution of (3.15) always yields a perfect recovery of ambiguous labeling such that $\mathbf{Y}^* = \mathbf{P}^0$. Several factors contribute to our inability to resolve the

ambiguity. For instance, if label 1 is consistently present in the candidate labeling set of each instance, assigning $\mathbf{v}_1$ for each column vector of $\mathbf{Y}$ yields a trivial solution. This issue is also addressed in [41], as learning from instances associated with two consistently co-occurring labels is impossible.

Note that $\mathbf{Y}^* = \mathbf{P}^0$ is one of the possible optimal solutions to (3.15). The solution may not be unique if any one of the instances belongs to more than one convex hull, i.e. the convex hulls from different classes overlap with each other. Hence, an instance can be ideally decomposed from either one of the convex hulls without further changing the rank of $[\mathbf{Y}; \mathbf{X}^0]$. This issue is analogous to the non-separable case of linear support vector machine (SVM). Nevertheless, it is our intention to seek $\mathbf{Y} = \mathbf{P}^0$ by solving (3.15) with the understanding that 1) the ambiguous labeling carries rational information, and 2) data lies in sufficiently high-dimensional space such that convex hulls of each class are separable [55].

Figure 3.2 illustrates the geometric interpretation of MCar using the convex hull representation. When each element in the candidate labeling set is trivially treated as the true label, the convex hulls of each class are erroneously expanded and the low-rank assumption of $[\mathbf{Y}; \mathbf{X}^0]$ does not hold. MCar exploits the underlying low-rank structure of $[\mathbf{Y}; \mathbf{X}^0]$, which is equivalent to reassigning the labels for those ambiguously labeled instances such that instances of the same class cohesively form a convex hull. Hence, each over-expanded convex hull shrinks to its actual contour, and the convex hulls become potentially separable. This is essentially different from discriminative ambiguous learning methods that construct the hyperplane between ambiguously labeled instances by minimizing the ambiguous loss.

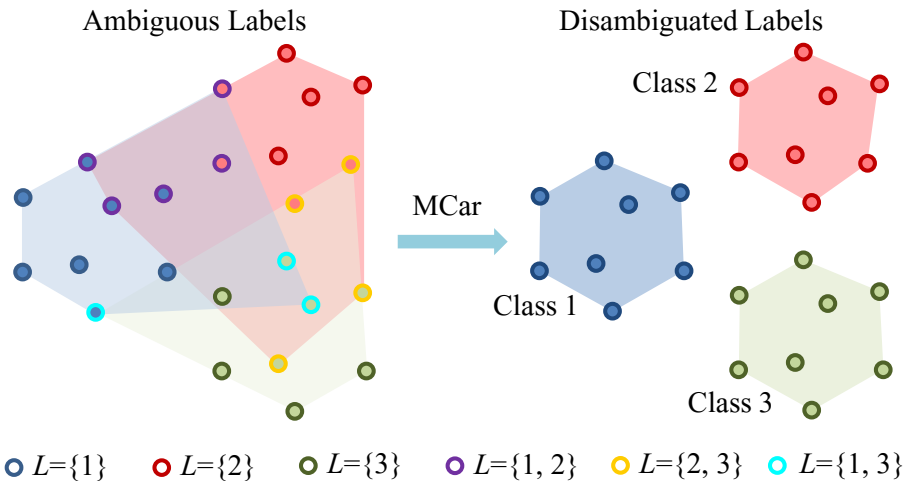


Figure 3.2: MCar reassigns the labels for those ambiguously labeled instances such that instances of the same subjects cohesively form potentially-separable convex hulls. The vertices of each convex hull are the representatives of each class, forming $\mathbf{D}_k$. The interior and outline of the circles are color-coded to represent three different classes and various ambiguous labels, respectively.

When data is contaminated by sparse errors, the optimization problem in (3.15) can be reformulated as

$$\begin{aligned} \min_{\mathbf{H},\mathbf{E}_X,\mathbf{E}_P} \ & \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 \\ \text{s.t. } \mathbf{H} = & \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix}, \\ & \mathbf{y}_j \in \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c\}, j = 1, 2, \ldots, N, \\ & y_{i,j} = 0 \text{ if } i \notin L_j \ \forall j, \end{aligned} \tag{3.16}$$

where $\mathbf{H}$ is the heterogeneous feature matrix in the absence of noise, and $\mathbf{Z}$ is the recovered feature matrix. The parameter $\lambda \in \mathbb{R}_+$ controls the rank of $\mathbf{H}$ and the sparsity of noise. The objective is to assign the predicted label $\mathbf{Y}$ and extract the sparse noise of $\mathbf{X}$ in pursuit of a low-rank $\mathbf{H}$. Figure 3.3 illustrates the ideal decomposition of the heterogeneous feature matrix, where the underlying low-rank structure and the ambiguous labels are recovered simultaneously.

As (3.16) is a combinatorial optimization problem, we relax each column vector of $\mathbf{Y}$ in probability simplex in $\mathbb{R}^c$. The original formulation can be rewritten as

$$\begin{aligned} \min_{\mathbf{H},\mathbf{E}_X,\mathbf{E}_P} \ & \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{Y}\|_0 \\ \text{s.t. } \mathbf{H} = & \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix}, \\ & \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \ \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \\ & y_{i,j} = 0 \text{ if } i \notin L_j \ \forall j, \end{aligned} \tag{3.17}$$

where $\gamma \in \mathbb{R}_+$ encourages the sparsity of $\mathbf{Y}$ such that the original discrete feasible region can be well approximated. From the perspective of convex hull representa-

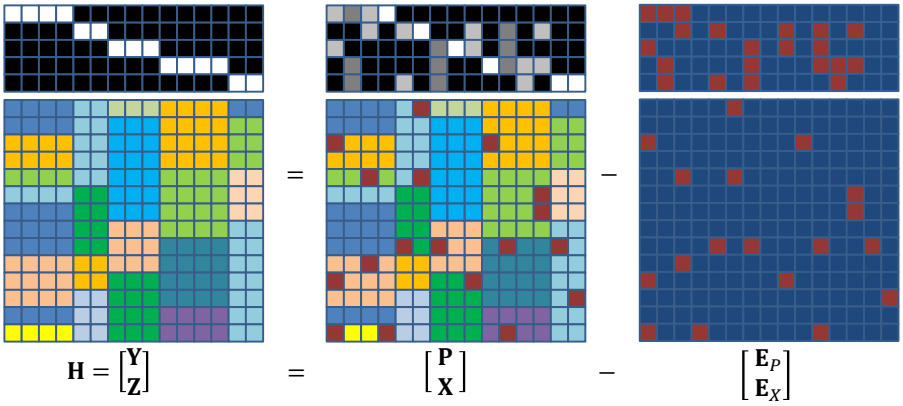


Figure 3.3: Ideal decomposition of the heterogeneous feature matrix using MCar. The underlying low-rank structure and the ambiguous labeling are recovered simultaneously.

tion, such relaxation allows each instance to be represented from more than one set of representative matrix $\mathbf{D}_k$, while it will be penalized by the non-sparsity of $\mathbf{Y}$. Consequently, the predicted label of instance j can be obtained as

$$\hat{l}_j = \arg\max_{i \in L_j} \; y_{i,j}. \tag{3.18}$$

3.2.3 Ambiguously Labeled Data with Labeling Imbalance

The class imbalance may lead to performance degradation in SVM as a majority class with abundant training samples can bias the decision boundary toward a minority class with scarce training samples. Analogously, MCar may suffer from labeling imbalance when a majority label is frequently present among the candidate labels in the ambiguously labeled data. When we resolve the ambiguity using (3.17), the heterogeneous feature vectors associated with a majority label are more likely to dominate the low-rank approximation of the heterogeneous matrix than those

associated with a minor label. Hence, the recovered soft labeling matrix will bias toward those soft labeling vectors associated with majority labels.

Class-weighted SVM applies unequal weighting to the cost function of different classes to mitigate the class imbalance [56]. Hence, instances from the minority label will be better emphasized than those from the dominant label to establish an objective decision boundary. However, the concept of class-weighted SVM cannot be directly applied to MCar to deal with label imbalances since each instance is not labeled as a particular class in the ambiguously labeled data. Without the knowledge of the true labels, we formulate the instance-weighted objective function of (3.14) as

$$\min_{\mathbf{T},\mathbf{D},\mathbf{Q}} \sum_{j=1}^{N} \eta_j \left\| \begin{bmatrix} \mathbf{p}_j \\ \\ \mathbf{x}_j \end{bmatrix} - \begin{bmatrix} \mathbf{T} \\ \\ \mathbf{D} \end{bmatrix} \mathbf{q}_j \right\|_F^2, \tag{3.19}$$

where η_j is the instance weight of the j^{th} instance. In order to balance the square errors contributed by each class in (3.19), we aim to set instance weight η_j as $1/N_{l_j}$, where N_{l_j} is the number of the instances from the l_j class. Nevertheless, assigning a class weight for each instance is not feasible in the ambiguously labeled data since the true label l_j is not explicitly known. Moreover, N_i is intractable since the data is not explicitly labeled. Hence, we propose to set the instance weight as

$$\eta_j = \frac{1}{\sum_{i=1}^{c} p_{i,j} \hat{N}_i}, \tag{3.20}$$

where

$$\hat{N}_i = \sum_{j=1}^{N} p_{i,j} \tag{3.21}$$

is the estimated number of instances of the i^{th} class. The estimated number of

instances of the i^{th} class accumulates the soft labeling scores corresponding to the i^{th} class across all the instances. With the soft labeling vector $\mathbf{p}_j$, we can compute the effective number of instances of the class that the j^{th} instance belongs to by $\sum_{i=1}^{c} p_{i,j}\hat{N}_i$. Hence, our proposed weighting scheme is eligible to compute the effective class weight of each ambiguously labeled instance even though the knowledge of true label is not available. The design of the instance weight is not unique, and readers may refer to [32,57] for modeling the instance weight with respect to various objectives.

For the ease of presentation, we reformulate (3.19) as

$$\min_{\mathbf{T},\mathbf{D},\mathbf{Q}} \left\| \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} \mathbf{W} - \begin{bmatrix} \mathbf{T} \\ \mathbf{D} \end{bmatrix} \mathbf{Q}\mathbf{W} \right\|_F^2, \tag{3.22}$$

where

$$\mathbf{W} = \sqrt{\mathrm{diag}(\mathbf{1}_N^T \mathbf{P}^T \mathbf{P})}^{-1} \tag{3.23}$$

is a diagonal weighting matrix with $w_{j,j} = \sqrt{\eta_j}$. As post-multiplying $\mathbf{W}$ does not increase the rank of a matrix, we claim that *Proposition 1* also applies to the weighted heterogeneous feature matrix $\mathbf{H}_{obs}\mathbf{W} = [\mathbf{P}; \mathbf{X}]\mathbf{W}$. We propose the weighted MCar

(WMCar) by generalizing (3.17) as

$$\begin{aligned}
&\min_{\mathbf{H},\mathbf{E}_X,\mathbf{E}_P} \operatorname{rank}(\mathbf{HW}) + \lambda\|\mathbf{E}_X\mathbf{W}\|_0 + \gamma\|\mathbf{YW}\|_0 \\
&\text{s.t. } \mathbf{HW} = \begin{bmatrix} \mathbf{YW} \\ \mathbf{ZW} \end{bmatrix} = \begin{bmatrix} \mathbf{PW} \\ \mathbf{XW} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P\mathbf{W} \\ \mathbf{E}_X\mathbf{W} \end{bmatrix}, \\
&\mathbf{1}_c^T\mathbf{YW} = \mathbf{1}_N^T\mathbf{W},\ \mathbf{YW} \in \mathbb{R}_+^{c\times N}, \\
&y_{i,j} = 0 \text{ if } i \notin L_j\ \forall j.
\end{aligned} \tag{3.24}$$

Let $\bar{\mathbf{H}}_{obs} = \mathbf{H}_{obs}\mathbf{W}$, $\bar{\mathbf{H}} = \mathbf{HW}$, and $\bar{\mathbf{E}} = \mathbf{EW}$, we reformulate (3.24) as

$$\begin{aligned}
&\min_{\bar{\mathbf{H}},\bar{\mathbf{E}}_P,\bar{\mathbf{E}}_X} \operatorname{rank}(\bar{\mathbf{H}}) + \lambda\|\bar{\mathbf{E}}_X\|_0 + \gamma\|\bar{\mathbf{Y}}\|_0 \\
&\text{s.t. } \bar{\mathbf{H}} = \begin{bmatrix} \bar{\mathbf{Y}} \\ \bar{\mathbf{Z}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{P}} \\ \bar{\mathbf{X}} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{E}}_P \\ \bar{\mathbf{E}}_X \end{bmatrix}, \\
&\mathbf{1}_c^T\bar{\mathbf{Y}} = \mathbf{1}_N^T\mathbf{W},\ \bar{\mathbf{Y}} \in \mathbb{R}_+^{c\times N}, \\
&\bar{y}_{i,j} = 0 \text{ if } i \notin L_j\ \forall j.
\end{aligned} \tag{3.25}$$

The predicted label can be retrieved from $\mathbf{Y} = \bar{\mathbf{Y}}\mathbf{W}^{-1}$ using (3.18). Interestingly, the instance-weighted MCar is equivalent to executing MCar with the weighted heterogeneous feature matrix. A larger weight on the heterogeneous feature vectors associated with minority labels provides those instances a stronger impact in the low-rank approximation of the heterogeneous matrix, and thus the labeling imbalance can be compensated. As (3.17) is generalized by (3.25) in consideration of labeling imbalance, WMCar is identical to MCar in the special case of $\mathbf{W} = \mathbf{I}$.

Algorithm 1 The optimization algorithm for WMCar (3.29)

Input: $\mathbf{P} \in \mathbb{R}^{c\times N}$, $\mathbf{X} \in \mathbb{R}^{m\times N}$, $\mathbf{W} \in \mathbb{R}^{N\times N}$, $L_j\ \forall j$, λ, and γ.

1: **Initialization:**

2: $\bar{\mathbf{P}} = \mathbf{PW}$, $\bar{\mathbf{X}} = \mathbf{XW}$, $\bar{\mathbf{H}}_{obs} = [\bar{\mathbf{P}}; \bar{\mathbf{X}}]$;

3: $\bar{\mathbf{Y}} = \mathbf{0}$, $\bar{\mathbf{Z}} = \mathbf{0}$, $\mu > 0, \mu_{\max} > 0$, $\rho > 1$, $\mathbf{\Lambda} = [\mathbf{\Lambda}_P; \mathbf{\Lambda}_X] = \bar{\mathbf{H}}_{obs}/\|\bar{\mathbf{H}}_{obs}\|_2$;

4: **while** not converged **do**

5: $\quad \bar{\mathbf{E}}_P = \bar{\mathbf{P}} - \mathcal{S}_{\gamma\mu^{-1}}[\bar{\mathbf{Y}} - \mu^{-1}\mathbf{\Lambda}_P]$;

6: $\quad \bar{\mathbf{E}}_X = \mathcal{S}_{\lambda\mu^{-1}}[\bar{\mathbf{X}} - \bar{\mathbf{Z}} + \mu^{-1}\mathbf{\Lambda}_X]$;

7: $\quad (\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}) = \text{svd}\left(\bar{\mathbf{H}}_{obs} - \bar{\mathbf{E}} + \mu^{-1}\mathbf{\Lambda}\right)$;

8: $\quad \bar{\mathbf{H}} = \mathbf{U}\mathcal{S}_{\mu^{-1}}[\mathbf{\Sigma}]\mathbf{V}^T$;

9: $\quad \mathbf{\Lambda} = \mathbf{\Lambda} + \mu\left(\bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}}\right)$;

10: $\quad \mu = \min(\rho\mu, \mu_{\max})$;

11: $\quad$ **Project** $\bar{\mathbf{Y}}$:

12: $\quad \triangleright$ Line: 13: Projection for (3.31)

13: $\quad \bar{y}_{i,j} = 0$ if $i \notin L_j\ \forall j$;

14: $\quad \triangleright$ Line: 15-16: Projection for (3.30)

15: $\quad \bar{\mathbf{Y}} = \max(\bar{\mathbf{Y}}, 0)$;

16: $\quad \bar{\mathbf{y}}_j = w_{j,j}\,\bar{\mathbf{y}}_j/\|\bar{\mathbf{y}}_j\|_1,\ \forall j$;

17: **end while**

18: $\mathbf{H} = \bar{\mathbf{H}}\mathbf{W}^{-1}$, $\mathbf{E} = \bar{\mathbf{E}}\mathbf{W}^{-1}$

Output: $(\mathbf{H}, \mathbf{E})$

3.3 Optimization

The augmented Lagrangian method (ALM) has been extensively used for solving low-rank problems [35, 58]. In this section, we propose to incorporate the ALM with the projection step [30, 31] to solve the optimization problem of WMCar.

In order to decouple $\bar{\mathbf{Y}}$ in the first and third terms of the objective function in (3.25), we replace $\|\bar{\mathbf{Y}}\|_0$ with $\|\bar{\mathbf{P}} - \bar{\mathbf{E}}_P\|_0$ and rewrite (3.25) as

$$\begin{aligned} \min_{\bar{\mathbf{H}},\bar{\mathbf{E}}_X,\bar{\mathbf{E}}_P} \ & \operatorname{rank}(\bar{\mathbf{H}}) + \lambda\|\bar{\mathbf{E}}_X\|_0 + \gamma\|\bar{\mathbf{P}} - \bar{\mathbf{E}}_P\|_0 \\ \text{s.t. } \bar{\mathbf{H}} &= \begin{bmatrix} \bar{\mathbf{Y}} \\ \bar{\mathbf{Z}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{P}} \\ \bar{\mathbf{X}} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{E}}_P \\ \bar{\mathbf{E}}_X \end{bmatrix}, \\ & \mathbf{1}_c^T\bar{\mathbf{Y}} = \mathbf{1}_N^T\mathbf{W}, \ \bar{\mathbf{Y}} \in \mathbb{R}_+^{c\times N}, \\ & \bar{y}_{i,j} = 0 \text{ if } i \notin L_j \ \forall j. \end{aligned} \tag{3.26}$$

Following the procedure of ALM, we relax the first constraint in (3.26) and reformulate it as

$$\begin{aligned} \min_{\bar{\mathbf{H}},\bar{\mathbf{E}},\mathbf{\Lambda},\mu} \ & \ell(\bar{\mathbf{H}}, \bar{\mathbf{E}}, \mathbf{\Lambda}, \mu) \\ \text{s.t. } & \mathbf{1}_c^T\bar{\mathbf{Y}} = \mathbf{1}_N^T\mathbf{W}, \ \bar{\mathbf{Y}} \in \mathbb{R}_+^{c\times N}, \\ & \bar{y}_{i,j} = 0 \text{ if } i \notin L_j \ \forall j, \end{aligned} \tag{3.27}$$

where $\mu \in \mathbb{R}_+$ and $\mathbf{\Lambda} \in \mathbb{R}^{(c+m)\times N}$. The Lagrangian is expressed as

$$\begin{aligned} \ell(\bar{\mathbf{H}}, \bar{\mathbf{E}}, \mathbf{\Lambda}, \mu) = & \operatorname{rank}(\bar{\mathbf{H}}) + \lambda\|\bar{\mathbf{E}}_X\|_0 + \gamma\|\bar{\mathbf{P}} - \bar{\mathbf{E}}_P\|_0 \\ & + \left\langle \mathbf{\Lambda}, \bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}} \right\rangle + \frac{\mu}{2}\left\|\bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}}\right\|_F^2. \end{aligned} \tag{3.28}$$

In order to make the optimization problem feasible, we approximate the rank with the nuclear norm and the ℓ_0 norm with the ℓ_1 norm [59]. Thus, we solve the following

formulation as the convex surrogate of (3.27)

$$\min_{\bar{\mathbf{H}},\bar{\mathbf{E}},\mathbf{\Lambda},\mu} \ \ell_R(\bar{\mathbf{H}},\bar{\mathbf{E}},\mathbf{\Lambda},\mu) \tag{3.29}$$

$$\text{s.t.} \ \ \mathbf{1}_c^T\bar{\mathbf{Y}} = \mathbf{1}_N^T\mathbf{W}, \ \ \bar{\mathbf{Y}} \in \mathbb{R}_+^{c\times N}, \tag{3.30}$$

$$\bar{y}_{i,j} = 0 \text{ if } i \notin L_j \ \forall j, \tag{3.31}$$

where the Lagrangian is represented as

$$\begin{aligned}\ell_R(\bar{\mathbf{H}},\bar{\mathbf{E}},\mathbf{\Lambda},\mu) = &\left\|\bar{\mathbf{H}}\right\|_* + \lambda\|\bar{\mathbf{E}}_X\|_1 + \gamma\|\bar{\mathbf{P}} - \bar{\mathbf{E}}_P\|_1 \\ &+ \left\langle \mathbf{\Lambda}, \bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}}\right\rangle + \frac{\mu}{2}\left\|\bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}}\right\|_F^2.\end{aligned} \tag{3.32}$$

The ALM operates in the sense that $\bar{\mathbf{H}}$, $\bar{\mathbf{E}}_P$, and $\bar{\mathbf{E}}_X$ can be solved alternately by fixing other variables. In each iteration, we employ a similar projection technique used in [30, 31] to enforce $\bar{\mathbf{Y}}$ to be feasible. The entire procedure for solving (3.29) is summarized in Algorithm 1, and the details of the optimization algorithm are presented in the following paragraphs.

3.3.1 Solving for $\bar{\mathbf{E}}_P$

To update $\bar{\mathbf{E}}_P$, we fix $\bar{\mathbf{H}}$, $\bar{\mathbf{E}}_X$, $\mathbf{\Lambda}$ and μ obtained in the previous iteration. Hence, the problem for updating $\bar{\mathbf{E}}_P$ can be solved by first computing

$$\begin{aligned}\bar{\mathbf{E}}_P^* = \underset{\bar{\mathbf{E}}_P}{\operatorname{argmin}} \ &\gamma\|\bar{\mathbf{P}} - \bar{\mathbf{E}}_P\|_1 + \left\langle \mathbf{\Lambda}_P, \bar{\mathbf{P}} - \bar{\mathbf{Y}} - \bar{\mathbf{E}}_P\right\rangle \\ &+ \frac{\mu}{2}\left\|\bar{\mathbf{P}} - \bar{\mathbf{Y}} - \bar{\mathbf{E}}_P\right\|_F^2.\end{aligned} \tag{3.33}$$

For the ease of derivation, we let $\bar{\mathbf{B}} = \bar{\mathbf{P}} - \bar{\mathbf{E}}_P$ and update $\bar{\mathbf{B}}$ as surrogate. We can reformulate (3.33) as

$$\begin{aligned}\bar{\mathbf{B}}^* &= \underset{\bar{\mathbf{B}}}{\operatorname{argmin}}\ \gamma\|\bar{\mathbf{B}}\|_1 + \left\langle \mathbf{\Lambda}_P, \bar{\mathbf{B}} - \bar{\mathbf{Y}} \right\rangle + \frac{\mu}{2}\left\|\bar{\mathbf{B}} - \bar{\mathbf{Y}}\right\|_F^2, \\ &= \underset{\bar{\mathbf{B}}}{\operatorname{argmin}}\ \gamma\|\bar{\mathbf{B}}\|_1 + \frac{\mu}{2}\|\bar{\mathbf{B}} - \bar{\mathbf{Y}} + \mu^{-1}\mathbf{\Lambda}_P\|_F^2, \\ &= \underset{\bar{\mathbf{B}}}{\operatorname{argmin}}\ \gamma\|\bar{\mathbf{B}}\|_1 + \frac{\mu}{2}\|\bar{\mathbf{Y}} - \mu^{-1}\mathbf{\Lambda}_P - \bar{\mathbf{B}}\|_F^2.\end{aligned} \tag{3.34}$$

Using the subgradient of (3.34), we can obtain the closed-form solution for updating $\mathbf{B}$

$$\bar{\mathbf{B}}^* = \mathcal{S}_{\gamma\mu^{-1}}[\bar{\mathbf{Y}} - \mu^{-1}\mathbf{\Lambda}_P]. \tag{3.35}$$

Consequently, we can update $\bar{\mathbf{E}}_P$ as

$$\bar{\mathbf{E}}_P^* = \bar{\mathbf{P}} - \bar{\mathbf{B}}^* = \bar{\mathbf{P}} - \mathcal{S}_{\gamma\mu^{-1}}[\bar{\mathbf{Y}} - \mu^{-1}\mathbf{\Lambda}_P]. \tag{3.36}$$

3.3.2 Solve $\bar{\mathbf{E}}_X$

To update $\bar{\mathbf{E}}_X$, we fix $\bar{\mathbf{H}}$, $\bar{\mathbf{E}}_P$, $\mathbf{\Lambda}$ and μ obtained in the previous iteration. Thus, the problem for updating $\bar{\mathbf{E}}_X$ can be solved by

$$\begin{aligned}\bar{\mathbf{E}}_X^* &= \underset{\bar{\mathbf{E}}_X}{\operatorname{argmin}}\ \lambda\|\bar{\mathbf{E}}_X\|_1 + \left\langle \mathbf{\Lambda}_X, \bar{\mathbf{X}} - \bar{\mathbf{Z}} - \bar{\mathbf{E}}_X \right\rangle \\ &\quad + \frac{\mu}{2}\left\|\bar{\mathbf{X}} - \bar{\mathbf{Z}} - \bar{\mathbf{E}}_X\right\|_F^2, \\ &= \underset{\bar{\mathbf{E}}_X}{\operatorname{argmin}}\ \lambda\|\bar{\mathbf{E}}_X\|_1 + \frac{\mu}{2}\|\bar{\mathbf{X}} - \bar{\mathbf{Z}} + \mu^{-1}\mathbf{\Lambda}_X - \bar{\mathbf{E}}_X\|_F^2.\end{aligned} \tag{3.37}$$

Using the subgradient of (3.37), we can obtain the closed-form solution for updating $\mathbf{E}_X$

$$\bar{\mathbf{E}}_X^* = \mathcal{S}_{\lambda\mu^{-1}}[\bar{\mathbf{X}} - \bar{\mathbf{Z}} + \mu^{-1}\mathbf{\Lambda}_X]. \tag{3.38}$$

3.3.3 Solve $\bar{\mathbf{H}}$

To update $\bar{\mathbf{H}}$, we fix $\bar{\mathbf{E}}_P$, $\bar{\mathbf{E}}_X$, $\mathbf{\Lambda}$ and μ obtained in the previous iteration. The feasible region of $\bar{\mathbf{Y}}$ in $\bar{\mathbf{H}}$ is currently not considered but will be handled in the projection step of $\bar{\mathbf{Y}}$ (Section 3.3.4). Therefore, the problem for updating $\bar{\mathbf{H}}$ can be solved by

$$\bar{\mathbf{H}}^* = \underset{\bar{\mathbf{H}}}{\operatorname{argmin}} \ \|\bar{\mathbf{H}}\|_* + \left\langle \mathbf{\Lambda}, \bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}} \right\rangle \tag{3.39}$$

$$+ \frac{\mu}{2} \left\| \bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}} \right\|_F^2, \tag{3.40}$$

$$= \underset{\bar{\mathbf{H}}}{\operatorname{argmin}} \ \|\bar{\mathbf{H}}\|_* + \frac{\mu}{2} \|\mathbf{A}_H - \bar{\mathbf{H}}\|_F^2, \tag{3.41}$$

where $\mathbf{A}_H = \bar{\mathbf{H}}_{obs} - \bar{\mathbf{E}} + \mu^{-1}\mathbf{\Lambda}$. According to [60], the above problem can be solved by

$$\bar{\mathbf{H}}^* = \mathbf{U}\mathcal{S}_{\mu^{-1}}[\mathbf{\Sigma}]\mathbf{V}^T, \tag{3.42}$$

where $\mathbf{\Sigma}$ can be obtained from the singular value decomposition (SVD) of $\mathbf{A}_H$ denoted as

$$(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}) = \operatorname{svd}(\mathbf{A}_H). \tag{3.43}$$

Following the procedure in the augmented Lagrangian method (ALM), we can update $\mathbf{\Lambda}$ and μ as

$$\mathbf{\Lambda} = \mathbf{\Lambda} + \mu \left(\bar{\mathbf{H}}_{obs} - \bar{\mathbf{H}} - \bar{\mathbf{E}} \right) \tag{3.44}$$

where

$$\mu = \min(\rho\mu, \mu_{\max}) \tag{3.45}$$

in each iteration based on the updated $\bar{\mathbf{E}}_P$, $\bar{\mathbf{E}}_X$, and $\bar{\mathbf{H}}$.

When the dimension of the heterogeneous feature matrix is large, computing the SVD of $\mathbf{A}_H$ in the singular value thresholding procedure is usually time-consuming. As an alternative, one can use the gradient ascend algorithm applied to the dual problem of (3.41) to solve $\bar{\mathbf{H}}$ [61].

3.3.4 Project $\bar{\mathbf{Y}}$

Since the SVD operation for solving $\bar{\mathbf{H}}$ does not always return a feasible $\bar{\mathbf{Y}}$, we use a projection technique similar to the one in [30, 31] to enforce $\bar{\mathbf{Y}}$ to be feasible in each iteration. The projection involves two steps. First, we enforce those entries of $\bar{\mathbf{Y}}$ that do not correspond to the candidate labels to be zeros since the actual label only comes from the candidate labeling set provided by the ambiguous labels. Second, each column vector of $\mathbf{Y} = \bar{\mathbf{Y}}\mathbf{W}^{-1}$ is constrained to be in the probability simplex. As a result, we replace those negative entries in $\bar{\mathbf{Y}}$ with zeros and then normalize each column $\bar{\mathbf{y}}_j$ so that the summation of the entries in $\bar{\mathbf{y}}_j$ is equal to $w_{j,j}$.

3.4 Iterative Candidate Elimination for Ambiguity Resolution

According to (3.23), the weighting matrix $\mathbf{W}$ of WMCar is a function of $\mathbf{P}$. As WMCar resolves the label ambiguity in $\mathbf{P}$, the recovered soft labeling matrix $\mathbf{Y}$ can provide a better estimate of $\mathbf{W}$ than the original $\mathbf{P}$. This motivates us to iteratively resolve the ambiguity by alternating between recovering $\mathbf{Y}$ and updating $\mathbf{W}$. Nevertheless, the performance of iterative WMCar is not steady as shown in

Algorithm 2 The algorithm for WMCar-ICE

Input: $\mathbf{P} \in \mathbb{R}^{c \times N}$, $\mathbf{X} \in \mathbb{R}^{m \times N}$, $L_j \ \forall j$.

1: **while** $\mathcal{A} \neq \varnothing$ and within the maximum number of iterations **do**

2: $\quad \mathbf{W} = \sqrt{\text{diag}(\mathbf{1}_N^T \mathbf{P}^T \mathbf{P})}^{-1}$;

3: $\quad$ Obtain $\mathbf{Y}$ using WMCar (Algorithm 1);

4: $\quad$ Eliminate the least likely candidate in L_j, $j \in \mathcal{E}$ using (3.46)-(3.49);

5: $\quad \triangleright$ Line: 6-7: Project $\mathbf{Y}$ to comply with $L_j, \forall j$

6: $\quad y_{i,j} = 0$, if $i \notin L_j \ \forall j$;

7: $\quad \mathbf{y}_j = \mathbf{y}_j / \|\mathbf{y}_j\|_1 \ \forall j$;

8: $\quad \mathbf{P} \leftarrow \mathbf{Y}$;

9: **end while**

Output: $(\mathbf{H}, \mathbf{E})$

Figure 3.15. We propose WMCar with ICE (WMCar-ICE) to resolve the ambiguity by WMCar and then remove the least likely candidate labels in each iteration. The least likely candidate label of the j^{th} instance is denoted as

$$m(j) = \underset{i \in L_j}{\operatorname{argmin}} \; y_{i,j}, \tag{3.46}$$

and its corresponding soft labeling score is denoted as $y_{m(j),j}$. As removing a candidate label, which is actually a true label, in the candidate set generates an irreversible error, we propose to iteratively remove a portion of the least likely candidate labels that have relatively low soft labeling scores than others.

Let $\mathcal{A}$ denote the set consisting of the indices of those instances that have more than one candidate label, which is represented as

$$\mathcal{A} = \{j \mid |L_j| > 1, \forall j\}. \tag{3.47}$$

We define the elimination factor as f_e $(0 \leq f_e \leq 1)$, which accounts for the proportion of instances in $\mathcal{A}$ participating in the candidate elimination. We construct a subset $\mathcal{E}$ of $\mathcal{A}$, which consists of the entries that correspond to the smallest f_e portion of $\{y_{m(j),j} | j \in \mathcal{A}\}$. We represent it as

$$\mathcal{E} = \{j \mid y_{m(j),j} \leq t, j \in \mathcal{A}\}. \tag{3.48}$$

Note that t is automatically determined such that $|\mathcal{E}| = \lceil f_e \, |\mathcal{A}| \rceil$. Hence, we can update the candidate labeling sets by

$$L_j \leftarrow L_j - \{m(j)\}, \; j \in \mathcal{E}. \tag{3.49}$$

We enforce the soft labeling matrix $\mathbf{Y}$ to comply with the updated candidate labeling sets. We set $y_{i,j} = 0$, if $i \notin L_j \; \forall j$ and project each column vector of $\mathbf{Y}$

in the probability simplex. The original $\mathbf{P}$ will be replaced by $\mathbf{Y}$, which will serve as the input of WMCar in the next iteration. The procedure of WMCar-ICE is summarized in Algorithm 2. Note that updating the weighting matrix $\mathbf{W}$ is an important step in WMCar-ICE since it adaptively adjusts the importance among instances based on the updated $\mathbf{Y}$ in the previous iteration. This ICE procedure can be utilized by other ambiguous learning techniques that adopt the soft labeling input/output similar to that of WMCar.

3.5 Labeling Constraints between Instances

In practical applications, several ambiguously labeled instances can appear in the same venue. As a result, pairwise relations between instances can be utilized to assist ambiguity resolution. For example, two persons in a news photo should not be identified as the same subject even though both of them are ambiguously labeled in the caption. Such prior knowledge can be easily incorporated by restricting the feasible region of the labeling matrix. Moreover, it is essential to handle the open set problem, where there are some instances whose identities never appear in the labels. These unrecognized instances can be treated as the null class.

In this section, we show how MCar's formulation can be extended to associate the identities in news photos when the names are provided in captions. We assume all the instances (face images) are collected from the K groups (photos), and G_k is the set of indices of the instances (face images) appearing in the k^{th} group (photo). Note that instances (face images) from the same group (photo) share the same

ambiguous labels provided by their associated caption. Without loss of generality, we assume that the c^{th} class corresponds to the null class. Considering the prior knowledge, the original formulation given in (3.17) can be reformulated as

$$\min_{\mathbf{H},\mathbf{E}_X,\mathbf{E}_P} \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{Y}\|_0 \tag{3.50}$$

$$\text{s.t. } \mathbf{H} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix},$$

$$\mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T,\ \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \tag{3.51}$$

$$y_{i,j} = 0 \text{ if } i \notin L_j,\ i = 1, 2, \ldots, c-1,\ \forall j, \tag{3.52}$$

$$\sum_{j\in G_k}\sum_{i=1}^{c-1} y_{i,j} \geq 1 \text{ if } \bigcup_{j\in G_k} L_j \neq \{c\}, \forall k, \tag{3.53}$$

$$\sum_{j\in G_k} y_{i,j} \leq 1,\ i = 1, 2, \ldots, c-1,\ \forall k. \tag{3.54}$$

Constraints (3.51) and (3.52) are inherited from the original formulation. The constraint in (3.53), assumes that there is at least one non-null identity in a photo unless all the instances in a photo are explicitly labeled as null. This constraint is enforced to avoid the trivial solution that all the instances are treated as belonging to the null class. A similar constraint has been considered by [38] and [33] via restricting the candidate labeling set and confining the feasible space of PPM, respectively. The constraint in (3.54) enforces the uniqueness of non-null identities. Note that this framework can be easily tailored to handle other prior knowledge (e.g. must/cannot-link constraints, prior statistics) by regularizing the labeling matrix. This problem can be solved by following the similar relaxation procedures for solving (3.17). The optimization procedure is summarized in Algorithm 3.

Following the relaxation procedure in Section 3.3, we can reformulate (3.50) as

$$\min_{\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu} \|\mathbf{H}\|_* + \lambda\|\mathbf{E}_X\|_1 + \gamma\|\mathbf{P} - \mathbf{E}_P\|_1 + \langle \mathbf{\Lambda}, \mathbf{H}_{obs} - \mathbf{H} - \mathbf{E}\rangle + \frac{\mu}{2}\|\mathbf{H}_{obs} - \mathbf{H} - \mathbf{E}\|_F^2, \tag{3.55}$$

$$\text{s.t. } \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T,\ \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \tag{3.56}$$

$$y_{i,j} = 0 \text{ if } i \notin L_j,\ i = 1, 2, \ldots, c-1,\ \forall j, \tag{3.57}$$

$$\sum_{j\in G_k}\sum_{i=1}^{c-1} y_{i,j} \geq 1 \text{ if } \bigcup_{j\in G_k} L_j \neq \{c\}, \forall k, \tag{3.58}$$

$$\sum_{j\in G_k} y_{i,j} \leq 1,\ i = 1, 2, \ldots, c-1,\ \forall k. \tag{3.59}$$

We use a similar procedure of Algorithm 1 presented in Section 3.3 to solve (3.55). We again use the projection method to guide the process of the matrix completion such that the constraints on $\mathbf{Y}$ are satisfied. Additionally, the projection of $\mathbf{Y}$ handles the group constraints such that the labeling constraints between instances are satisfied. Hence, we project $\mathbf{Y}$ to the feasible regions indicated by (3.57), (3.58), and (3.59) one at a time, and each one is followed by the projection onto the feasible region indicated by (3.56) to ensure that each column of $\mathbf{Y}$ lies in the probability simplex. The detailed procedure is summarized in Algorithm 3. This algorithm can be easily extended to handle ambiguously labeled data with labeling imbalance by taking $\bar{\mathbf{H}}$ as input with proper manipulation on the projection steps of $\bar{\mathbf{Y}}$.

Algorithm 3 The optimization algorithm for (3.55)

Input: $\mathbf{P} \in \mathbb{R}^{c\times N}$, $\mathbf{X} \in \mathbb{R}^{m\times N}$, L_j $\forall j$, G_k $\forall k$, λ, and γ.

1: **Initialization:** $\mathbf{Y}=\mathbf{0}$, $\mathbf{Z}=\mathbf{0}$, $\mu>0, \mu_{\max}>0$, $\rho>1$, $\mathbf{\Lambda}=[\mathbf{\Lambda}_P;\mathbf{\Lambda}_X]=\mathbf{H}_{obs}/\|\mathbf{H}_{obs}\|_2$;
2: **while** not converged **do**
3: $\mathbf{E}_P=\mathbf{P}-\mathcal{S}_{\gamma\mu^{-1}}[\mathbf{Y}-\mu^{-1}\mathbf{\Lambda}_P]$;
4: $\mathbf{E}_X=\mathcal{S}_{\lambda\mu^{-1}}[\mathbf{X}-\mathbf{Z}+\mu^{-1}\mathbf{\Lambda}_X]$;
5: $(\mathbf{U},\mathbf{\Sigma},\mathbf{V})=\text{svd}\left(\mathbf{H}_{obs}-\mathbf{E}+\mu^{-1}\mathbf{\Lambda}\right)$;
6: $\mathbf{H}=\mathbf{U}\mathcal{S}_{\mu^{-1}}[\mathbf{\Sigma}]\mathbf{V}^T$;
7: $\mathbf{\Lambda}=\mathbf{\Lambda}+\mu\left(\mathbf{H}_{obs}-\mathbf{H}-\mathbf{E}\right)$;
8: $\mu=\min(\rho\mu,\mu_{\max})$;
9: **Project** $\mathbf{Y}$:
10: ▷ Line: 11-13: Projection for (3.57) and (3.56)
11: $\mathbf{Y}=\max(\mathbf{Y},0)$;
12: $y_{i,j}=0$ if $i\notin L_j,\ i=1,2,\ldots,c-1,\forall j$;
13: $\mathbf{y}_j=\mathbf{y}_j/\|\mathbf{y}_j\|_1,\ \forall j$;
14: ▷ Line: 15-22: Projection for (3.58) and (3.56)
15: **for** $k=1:K$ **do**
16: **if** $\cup_{j\in G_k}L_j\neq\{c\}$ **then**
17: **for** $i=1:c-1, j\in G_k$ **do**
18: $y_{i,j}=y_{i,j}/\min(\sum_{g\in G_k}\sum_{i=1}^{c-1}y_{i,g},1)$;
19: **end for**
20: **end if**
21: **end for**
22: $\mathbf{y}_j=\mathbf{y}_j/\|\mathbf{y}_j\|_1,\ \forall j$;
23: ▷ Line: 24-29: Projection for (3.59) and (3.56)
24: **for** $k=1:K$ **do**
25: **for** $i=1:c-1, j\in G_k$ **do**
26: $y_{i,j}=y_{i,j}/\max(\sum_{g\in G_k}y_{i,g},1)$;
27: **end for**
28: **end for**
29: $\mathbf{y}_j=\mathbf{y}_j/\|\mathbf{y}_j\|_1,\ \forall j$;
30: **end while**

Output: $(\mathbf{H},\mathbf{E})$

3.6 Experimental Results

We use the Labeled Faces in the Wild (LFW) dataset [62] and the CMU PIE dataset with synthesized ambiguous labels to evaluate the performance of our method under various controlled parameter settings. Furthermore, we use the *Lost* dataset [10] and the Labeled Yahoo! News dataset [24, 63] to demonstrate the effectiveness of our method in real-world applications. For datasets provided with face images, we use face images in gray scale of range $[0, 1.0]$. Each instance is preprocessed with histogram equalization and converted into a column feature vector.

3.6.1 Parameters

It is interesting to observe that (3.16) becomes asymptotically similar to the formulation of Robust Principle Component Analysis (RPCA) [35] as the dimension of the data feature is far greater than the number of classes. Motivated by this fact, we fix λ as

$$\lambda_o = \frac{1}{\sqrt{\max(c+m, N)}}, \tag{3.60}$$

which is the tradeoff parameter suggested in RPCA. γ is a tuning parameter that controls the sparsity of the soft labeling vectors. For MCar-based methods, we use $\gamma = 2\lambda_0$ to encourage stronger sparsity of the labeling vector than that of feature noise. For ICE, we set the elimination factor f_e as 0.5, and set the maximum number of iterations as 5. These parameters yield good results in general, and we will investigate the sensitivity of parameters in Section 3.6.4.

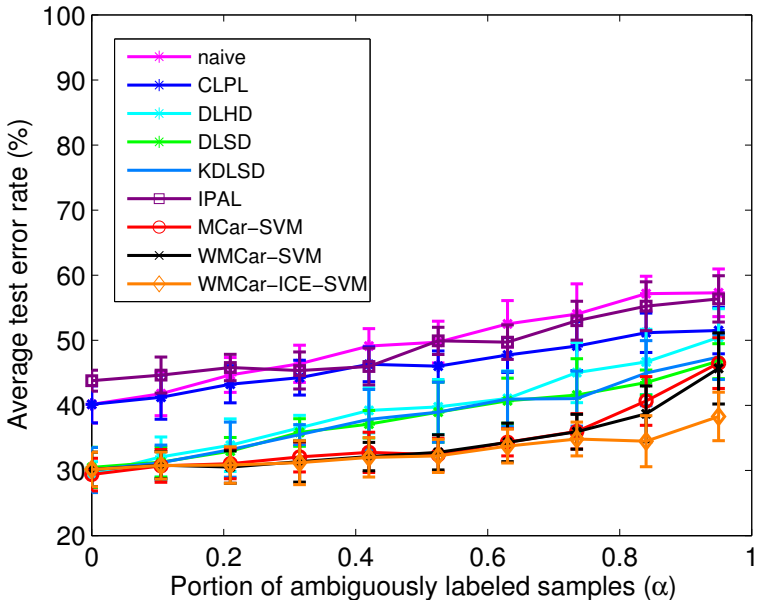


Figure 3.4: Performance comparisons on the FIW(10b) dataset. $\alpha \in [0, 0.95]$, $\beta = 2$, *inductive* experiment.

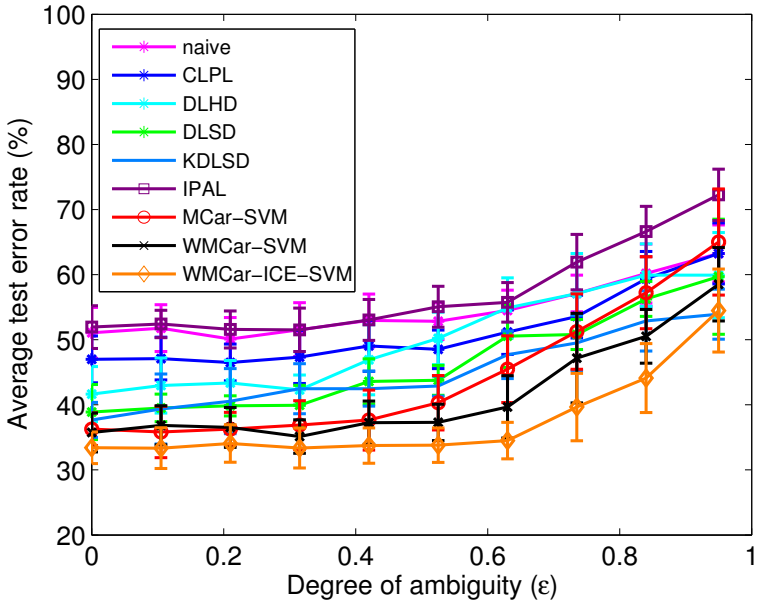


Figure 3.5: Performance comparisons on the FIW(10b) dataset. $\alpha = 1.0$, $\beta = 1$, $\epsilon \in [1/(c-1), 1]$, *inductive* experiment.

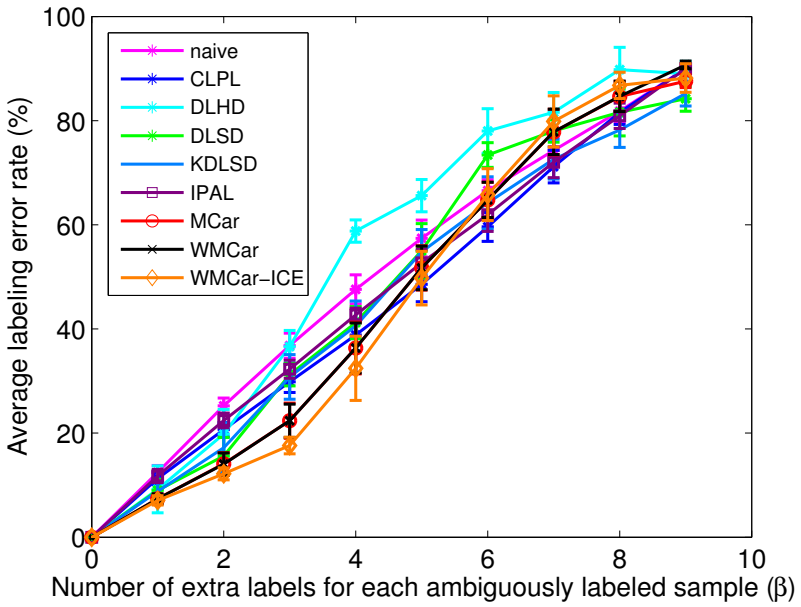


Figure 3.6: Performance comparisons on the FIW(10b) dataset. $\alpha = 1.0$, $\beta \in [0, 1, \ldots, 9]$, *transductive* experiment.

3.6.2 Experiments with the Synthesized Datasets

We conduct two types of controlled experiments suggested in [41]. For the *inductive* experiment, the dataset is evenly split into ambiguously labeled training set and unlabeled testing set. The proposed methods, MCar/WMCar-SVM and WMCar-ICE-SVM, learn a multi-class linear SVM [18] with the disambiguated labels provided by MCar/WMCar and WMCar-ICE, respectively. The testing data is then classified using the learned classifier. For the *transductive* experiment, all the data is used as the ambiguously labeled training set.

We follow the ambiguity model defined in [41] to generate ambiguous labels in the controlled experiment. Note that α denotes the number of extra labels for each instance, and β represents the portion of the ambiguously labeled data among

all the instances. The degree of ambiguity ϵ indicates the maximum probability that an extra label co-occurs with a true label, over all labels and instances. Each controlled experiment is repeated 20 times. We report the average testing (labeling) error rate for inductive (transductive) experiment, where the testing (labeling) error rate is the ratio of the number of erroneously labeled instances to the total number of instances in the testing (training) set. The standard deviations are plotted as error bars in the figures.

We compare the proposed MCar-based methods with several state-of-the-art ambiguous learning approaches for single instances with ambiguous labeling: CLPL [41], DLHD/DLSD [28], KDLSD [34], and IPAL [45]. We report the performance of these methods when the experimental results are available in their papers. Otherwise, we use the configuration suggested in their papers to conduct the experiments. We use 'naive' [41] as the baseline method, which learns a classifier from minimizing the trivial 0/1 loss.

3.6.2.1 The LFW Dataset

The FIW(10b) dataset [41] consists of the top 10 most frequent subjects selected from the LFW dataset [62], and the first 50 face images of each subject are used for evaluation. We use the cropped and resized face images readily provided by the authors of [41], where the face images are of 45×55 pixels.

Figures 3.4 and 3.5 show the results of the inductive experiments. Figure 3.4 shows that the MCar-based methods significantly outperform all the other methods

when the portion of ambiguously labeled data is larger than 0.2. The performance of WMCar is comparable to that of MCar since the ambiguously labeled data generated by this ambiguity model does not substantially result in labeling imbalance. WMCar-ICE demonstrates better performance than MCar and WMCar when more than 0.7 portions of the instances are ambiguously labeled. An explanation is that ICE eliminates the candidates based on the ordering of the least soft labeling score of each instance. This prioritization step can effectively benefit from a large portion of ambiguously labeled samples (i.e., large α) that usually carries a diverse aspect of soft labeling scores. When the portion of ambiguously labeled samples is small, the improvement due to ICE becomes insignificant.

Figure 3.5 shows that MCar outperforms prior methods over various degrees of ambiguity except when $\epsilon > 0.7$. Thus, MCar yields improved performance at low and intermediate levels of ambiguity, but it becomes susceptible at high levels of ambiguity. One explanation is that both the true label and the extra labels of a subject will result in low-rank component of the labeling matrix when they are likely to co-occur in high degree of ambiguity. Consequently, separating the true label from the extra labels in MCar becomes challenging. Another explanation is that a high degree of ambiguity results in labeling imbalance, which causes the performance degradation of MCar. To verify this, we obtain the label distribution by counting the number of label occurrences in the candidate labeling sets for each class. We define the imbalance factor as the ratio of the maximum to the minimum value in the label distribution. The average imbalance factor varies from 1.33 to 3.58 as the degree of ambiguity increases. This confirms that WMCar outperforms MCar in high

degree of ambiguity since WMCar is effective in mitigating the impact of labeling imbalance. Furthermore, WMCar-ICE outperforms WMCar by iteratively removing the least likely candidate labels from the candidate labeling sets. This experiment demonstrates that the labeling imbalance can cause performance degradation even though there is no class imbalance among the number of groundtruth faces per class.

In Figure 3.6, MCar-based methods outperform the other approaches only when the number of extra labels is less than 5 in the transductive experiment. This shows that MCar-based methods cannot be effective when the labeling is severely cluttered such that the low-rank approximation of heterogeneous feature fails. Similar to the controlled parameter setting in Figure 3.4, the ambiguously labeled data generated by this ambiguity model does not substantially result in labeling imbalance. Hence, the performance of WMCar is comparable to that of MCar, and WMCar-ICE slightly outperforms WMCar.

Figure 3.9 shows the intermediate results of low-rank decomposition of the feature matrix using MCar. Note that variations due to illumination, occlusions (e.g. eyeglasses, hand), and expressions are suppressed such that the low-rank component of a subject is preserved. In contrast to MCar-based methods, the discriminative methods (e.g. naive, CLPL) and IPAL are susceptible to such variations. Furthermore, it also demonstrates the robustness of our methods even though the face images are not perfectly aligned. The proposed method outperforms the dictionary-based methods [28, 34] for all cases except when there is severe ambiguity. Note the low-rank approximation of MCar operates on the feature matrix and ambiguous labeling matrix as a whole by concatenating them such that the actual labels and

the low-rank component of feature matrix are recovered simultaneously. This essentially demonstrates the advantage of the proposed method over the DLHD/DLSD and KDLSD methods that iteratively alternate between confidence and dictionary update.

3.6.2.2 The CMU PIE Dataset

The CMU PIE dataset contains face images from 68 subjects of different poses, illumination conditions, and expressions. Following the protocol presented in [28], we select the 18 subjects for evaluation. Each subject has 21 images under different illumination conditions, and the face images are resized to 40×48 pixels.

We synthesize the ambiguous labels based on the controlled parameters. The results of two transductive experiments for CMU PIE dataset are shown in Figures 3.7 and 3.8. In Figure 3.7, MCar-based methods and IPAL recover all the label ambiguity for various portions of ambiguously labeled samples. In Figure 3.8, our proposed methods consistently outperform most of the state-of-the-art methods except IPAL as we increase the number of extra labels for each ambiguously labeled sample. Since the CMU PIE dataset is collected in a constrained environment, the collective face images of a subject are well-modeled by low-rank approximation. Hence, MCar demonstrates marginally improvements over most of the methods in this dataset. This can be seen by visualizing the intermediate results of low-rank decomposition of the feature matrix using MCar as shown in Figure 3.10. Besides, the IPAL method outperforms our methods when $\beta > 6$. Since the IPAL method

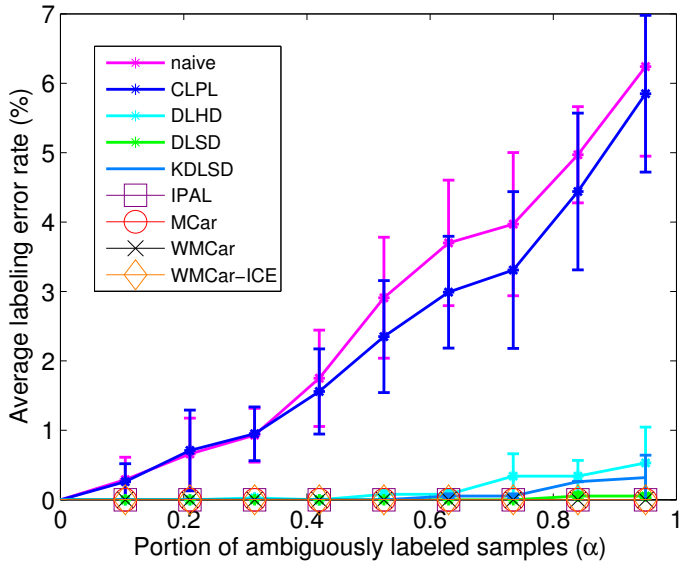


Figure 3.7: Performance comparisons on the CMU PIE dataset. $\alpha \in [0, 0.95]$, $\beta = 2$, *inductive* experiment.

utilizes the locally linear embedding for label propagation, which is effective in learning the underlying structure of data that has plenty of samples collected in the constrained environment. Hence, IPAL is able to recover the severely cluttered labels that MCar-based methods fail to approximate it as a low-rank matrix.

3.6.3 Experiments with Real-world Datasets

We conduct experiments on the *Lost* dataset and Labeled Yahoo! News dataset where the ambiguous labeling are collected in the real world. In the Labeled Yahoo! News dataset, we consider the labeling constraints between instances.

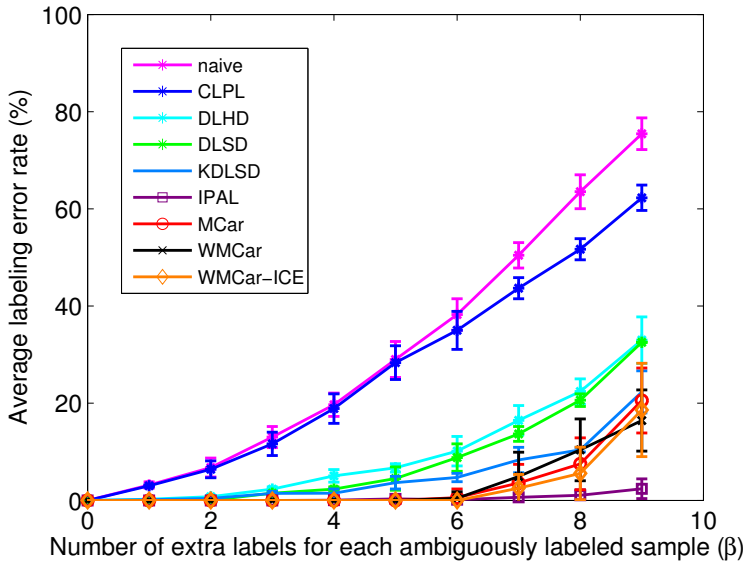


Figure 3.8: Performance comparisons on the CMU PIE dataset. $\alpha = 1.0$, $\beta \in [0, 1, \ldots, 9]$, *transductive* experiment.

Figure 3.9: A subset of images from FIW(10b) demonstrates the low-rank decomposition of feature matrix in MCar: the original face images, histogram-equalized images $\mathbf{X}$, low-rank component $\mathbf{Z}$, and noisy component $\mathbf{E}_X$, from the first row to the forth row, respectively.

Figure 3.10: A subset of images from the CMU PIE dataset demonstrates the low-rank decomposition of feature matrix in MCar: the original face images, histogram-equalized images $\mathbf{X}$, low-rank component $\mathbf{Z}$, and noisy component $\mathbf{E}_X$, from the first row to the forth row, respectively.

3.6.3.1 The Lost Dataset

The *Lost* dataset consists of face images and ambiguous labels automatically extracted using the screenplays provided in the TV series *Lost*. We use the *Lost* $(16, 8)$ dataset released by the authors of [10] for evaluation. The *Lost* $(16, 8)$ dataset consists of 1122 registered face images from 8 episodes, and the size of each is 60×90 pixels. The labels cover 16 subjects, but only 14 of them appear in the dataset. Figure 3.11 illustrates the label distribution, which exhibits labeling imbalance.

We compare MCar-based methods with the performance of 'naive', CLPL , MMS [38], and IPAL [45]. No labeling constraint between instances is considered in this experiment. Results are shown in Table 3.1. It can be seen from this table that MCar-based methods significantly outperform CLPL and MMS. This shows that

MCar-based methods resolve the ambiguity and handles variations of instances in the TV series much better when compared to discriminative methods. Note that the performance of MMS is close to that of CLPL since the ambiguous loss functions of both methods become similar when no labeling constraint between the instances is considered.

Figure 3.12 demonstrates that the groundtruth label distribution estimated by (3.21) is close to the groundtruth. Hence, WMCar can effectively utilize this information to compensate the labeling imbalance. Although WMCar slightly outperforms MCar, the collaboration of WMCar and ICE (WMCar-ICE) significantly outperforms MCar. On the other hand, MCar-ICE is inferior to WMCar-ICE since the ICE procedure can inadvertently remove the candidates corresponding to minor labels without considering the labeling imbalance. It is challenging for IPAL to exploit the underlying structure of scarcity labeled data and deal with labeling imbalance. Hence, IPAL cannot successfully resolve the label ambiguity in this dataset. We tailor the RPCA [58] and MC-Pos [54] to solve the ambiguous learning problem by trivially taking the heterogeneous matrix as input and predicting the labels from the soft labeling matrix of output with (3.18). The experimental result shows that existing low-rank approximation methods cannot substantially resolve the label ambiguity.

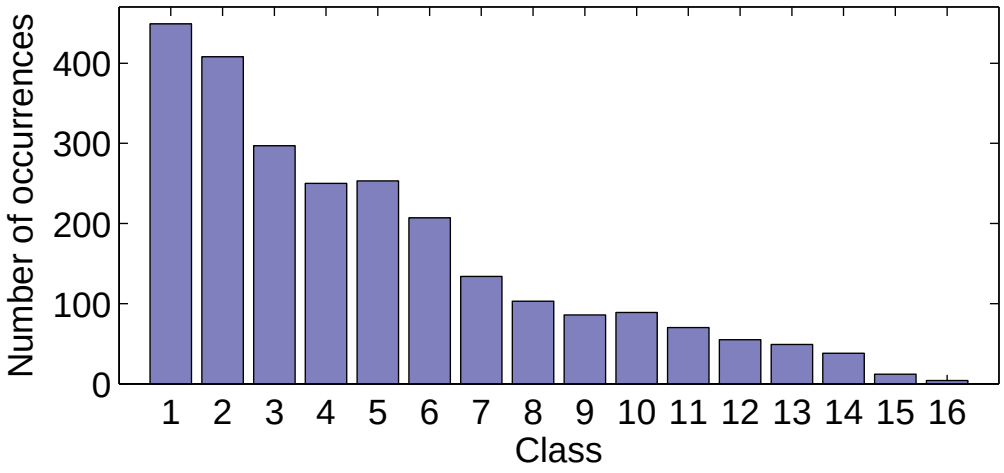


Figure 3.11: The label distribution of the *Lost* $(16, 8)$ dataset.

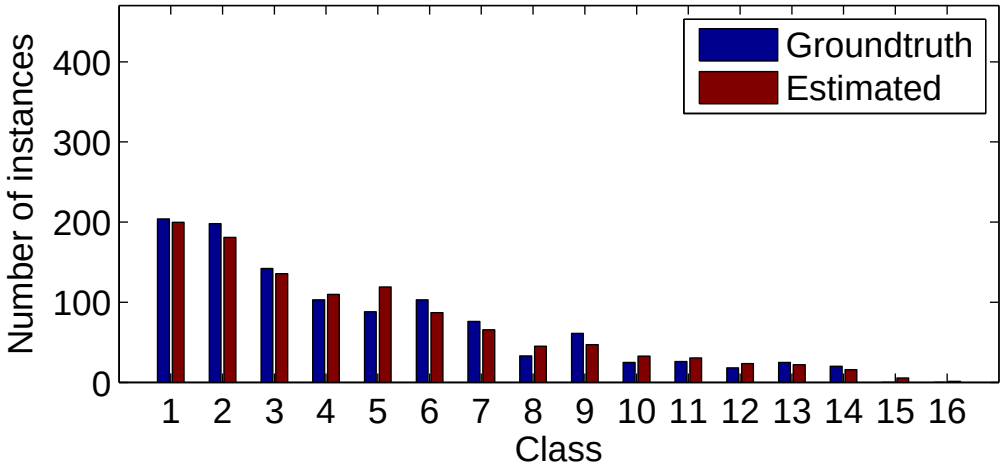


Figure 3.12: The groundtruth label distribution of the *Lost* $(16, 8)$ dataset. 'Groundtruth' denotes the number of instances per class counted from the groundtruth labels, and 'Estimated' denotes the estimate of the groundtruth label distribution from the ambiguous labels.

Table 3.1: Labeling error rates for the *Lost* $(16, 8)$ dataset (available at http://www.timotheecour.com/tv_data/tv_data.html).

Method	**Error Rate**
naive	18.6 %
CLPL [41]	12.6 %
MMS [38]	11.4 %
IPAL [45]	22.9 %
RPCA [58]	29.9 %
MC-Pos [54]	23.6 %
MCar	8.5 %
WMCar	8.2 %
MCar-ICE	8.0 %
WMCar-ICE	5.2 %

Table 3.2: Average testing error rates for the Labeled Yahoo! News dataset (available at http://lear.inrialpes.fr/data).

Method	**Error Rate**
CL-SVM	23.1 % ± 0.6 %
MIMLSVM [64]	25.3 % ± 0.3 %
MMS [38]	14.3 % ± 0.5 %
LR-SVM [33]	19.2 % ± 0.4 %
MCar-SVM	14.5 % ± 0.4 %
WMCar-SVM	13.6 % ± 0.8 %
MCar-ICE-SVM	15.0 % ± 1.0 %
WMCar-ICE-SVM	12.9 % ± 0.8 %

3.6.3.2 The Labeled Yahoo! News Dataset

The Labeled Yahoo! News dataset contains fully annotated faces in the images with names in the captions. It consists of 31147 detected faces from 20071 images. We use the precomputed SIFT feature of dimension 4992 extracted from that face images provided by Guillaumin *et al.* [63]. Following the protocol suggested in [38], we retain the 214 subjects with at least 20 occurrences in the captions. The remaining face images and names are treated as belonging to the additional null class. The ambiguous labeling is imbalanced in this dataset, where the number of labels present in the captions ranges from 20 to 1917 with mean and standard deviations equal to 64.6 and 147.3, respectively. The top two subjects that are present most frequently in the captions are 'george_w_bush' and 'saddam_hussein'. We conduct experiments on five training/testing splits by randomly selecting 80% of images and their associated captions as training set, and the rest are used as testing set. In each split, we also maintain the ratio between the number of training and testing instances from each subject.

The baseline approaches are CL-SVM and MIMLSVM [64], where their implementation details are provided in [38]. We compare with two state-of-the-art ambiguous labeling methods that consider labeling constraints between instances: MMS [38] and LR-SVM [33], which are based on discriminative model and low-rank framework, respectively. We resolve the ambiguity for the labels in the training set using (3.50) and train a multi-class linear SVM [18] to classify the testing data. Our MCar-SVM algorithm exhibits a slightly 0.2% higher error rate as compared

to MMS. An explanation is that MCar relying on the low-rank approximation for ambiguity resolution is particularly sensitive to labeling imbalance. This results in performance degradation in the learned classifier since the output labels of MCar are potentially biased toward the majority labels.

Compared to the LR-SVM method, the MCar-SVM algorithm demonstrates 4.7% improvement on the testing accuracy. Since MCar assigns the labels across all instances via low-rank approximation of heterogeneous feature matrix, it is more effective than the LR-SVM method, which updates the PPM and the low-rank subspace of each class alternately. When we consider the labeling imbalance and utilize the ICE procedure, our proposed WMCar-ICE-SVM outperforms MCar-SVM by 1.6%

3.6.4 Sensitivity of Parameters

We use the *Lost* $(16, 8)$ dataset to conduct the sensitivity analysis of MCar-based methods. In Figure 3.13, we evaluate the performance of WMCar over a set of parameters (λ, γ). We observe that the labeling error rate is relatively low when λ approaches λ_0 with respect to various γ. A similar trend is also observed when we evaluate the performance of MCar with the same experimental setting. Hence, we conclude that the tradeoff parameter suggested in RPCA is applicable or at least a good reference for selecting λ.

In Figure 3.14, we evaluate the performance of MCar-based methods with various γ and a fixed $\lambda = \lambda_o$. For ICE, we set the elimination factor f_e as 0.5,

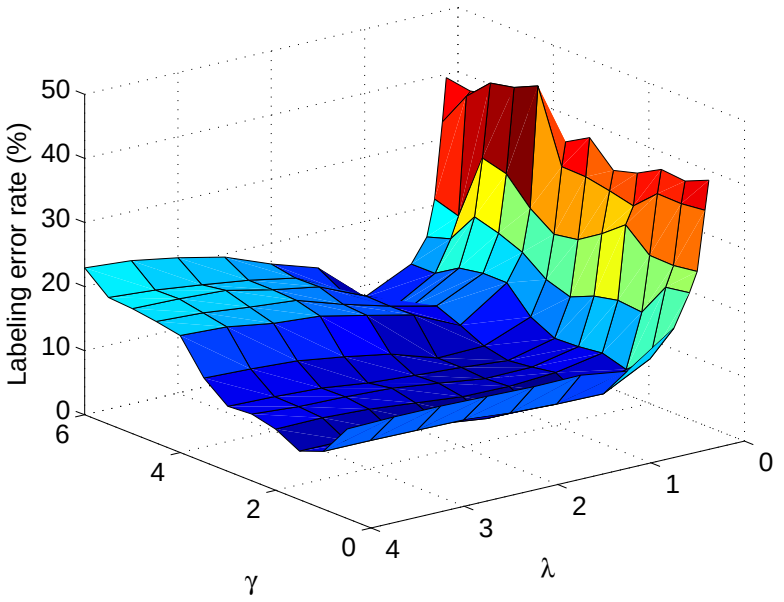


Figure 3.13: Labeling error rates of WMCar evaluated with a set of parameters (λ, γ) in the *Lost* $(16, 8)$ dataset. The λ-axis and γ-axis are normalized with respect to λ_o.

and we set the maximum number of iterations as 5. It shows that WMCar outperforms MCar for $\gamma \in [1, 4]\lambda_o$. WMCar-ICE outperforms WMCar and MCar for $\gamma \in [0.25, 6]\lambda_o$. We empirically set $\gamma = 2\lambda_o$, which yields good results for the MCar-based methods as illustrated in Figure 3.14. The performance of MCar-ICE flattens out as $\gamma \in [1.5, 6]\lambda_o$, which is inferior to WMCar-ICE. This again confirms that WMCar is essential to ICE to effectively reduce the labeling error. Besides, we observe that WMCar-ICE is less sensitive to γ since the ICE procedure intrinsically encourages the sparsity when removing the least likely candidate from a candidate labeling set. It is interesting to note that the lowest labeling error rate 3.7% is attained by WMCar-ICE with $\gamma = 1.75\lambda_o$. With the availability of validation data, a properly-selected γ can yield remarkable performance of MCar-based methods.

We conduct the sensitivity analysis of WMCar-ICE with $\lambda = \lambda_o$ and $\gamma = 2\lambda_o$ and evaluate the performance with various f_e. In Figure 3.15, the performance of WMCar-ICE ($f_e = 0$) fluctuates since the ICE procedure becomes ineffective as $f_e = 0$. A small elimination factor ($f_e = 0.25$) yields better performance than large elimination factors, but it takes more iterations to converge. Since the candidate elimination step in WMCar-ICE can incur an irreversible error, a small elimination factor can conservatively eliminate the least likely candidates in the candidate labeling sets. Hence, abrupt decision resulting from large elimination factors can be avoided, and the soft labeling matrix can be gently updated to guide the low-rank approximation of heterogeneous matrix. Considering the tradeoff between the rate of convergence and performance, we set $f_e = 0.5$ and let the maximum number of iterations be 5 in ICE.

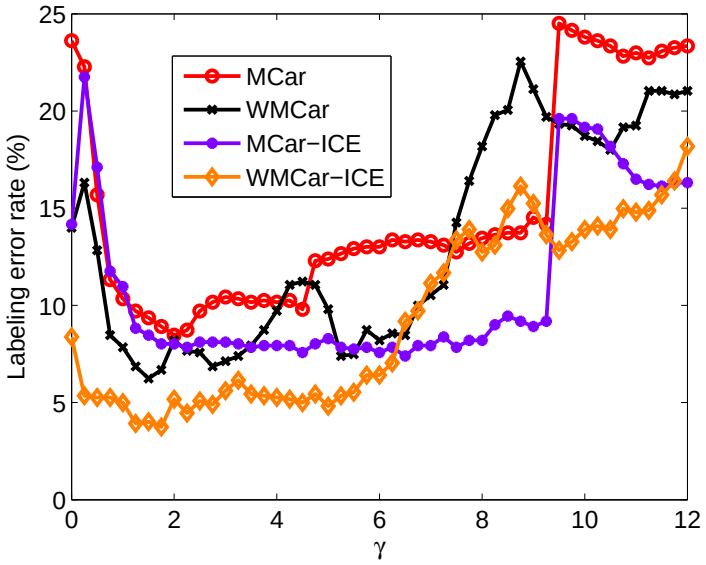


Figure 3.14: Labeling error rates of MCar-based methods versus γ in the *Lost* $(16, 8)$ dataset with $\lambda = \lambda_o$. The γ-axis is normalized with respect to λ_o.

3.6.5 Convergence

Since the projection method in Section 3.3.4 is not non-expansive, we cannot simply follow the rationale that the composition of gradient, shrinkage, and projection steps is non-expansive to prove convergence [54]. We attempt to replace the projection method of MCar with the Euclidean projection onto the simplex [65], which is a non-expansive projection, but the performance of the modified MCar degrades significantly. An explanation is that the Euclidean projection onto the simplex can inadvertently generate non-sparse entries, which conflicts with the original objective to encourage the sparsity of the soft labeling matrix in (3.25). On the other hand, our simple projection step normalizes the ℓ_1 norm of a soft labeling vector, which effectively restricts the soft labeling vector to lying on the ℓ_1 ball

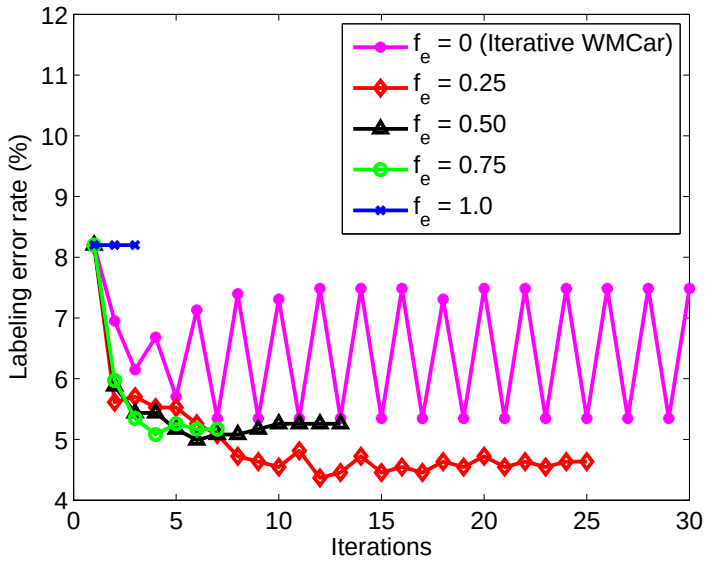


Figure 3.15: Labeling error rate versus the number of iterations in WMCar-ICE. The performance is evaluated in the *Lost* $(16, 8)$ dataset with various elimination factors. The performance of WMCar-ICE ($f_e = 0$) fluctuates since the ICE procedure becomes ineffective as $f_e = 0$.

and maintains an identical sparsity. Although the convergence of MCar has been observed empirically in [66], a theoretical justification of convergence needs further investigation. Since the number of ambiguous labels is finite, the convergence of ICE is straightforward with $f_e > 0$

3.7 Summary

We introduced a novel matrix completion framework for resolving the ambiguity of labels. In contrast to existing iterative alternating approaches, the proposed MCar method ensures all the instances and their associated ambiguous labels are utilized as a whole for resolving the ambiguity. Since MCar is capable of discovering the underlying low-rank structure of subjects, it is robust to within-subject variations. Hence, MCar can serve as the counterpart of discriminative ambiguous learning methods. Besides, WMCar generalizes MCar to compensate the labeling imbalance, and thus an instance associated with minority labels has a stronger impact than that associated with majority labels. The ICE procedure improves the performance of iterative WMCar by eliminating a portion of the least likely candidates in each iteration. As demonstrated by the experiments on the synthesized ambiguous labels and two datasets collected from real world, our proposed methods consistently resolve the ambiguity when single face images or group of face images are ambiguously labeled.

Chapter 4: Video-based Face Association and Identification

Video-based face identification [67–69] has broad applications, such as automatic indexing of a video, shot retrieval of a character in a TV-series, and suspect identification in surveillance videos. Unlike still images, a subject in a video generates diverse exemplars that contribute to creating a robust representation. Videos of these applications usually consist of multiple shots that involve scene and view changes. Nevertheless, most of the current video identification techniques focus on the identification task where the videos are of a single shot and the frame-by-frame face bounding boxes of the target (i.e., person of interest) are either readily provided or automatically associated using a tracking algorithm. For instance, the YouTube Faces dataset [70] provides frame-by-frame annotations, which has been used as benchmarks for evaluating video-based face identification algorithm. Although bounding boxes can be automatically extracted with face detection and tracking, human supervision is often needed to ensure that the annotations are not corrupted by the failure of face detection and tracking steps. Besides, most video-based identification techniques [69, 71–73] have evaluated their performances on video face datasets consisting of single-shot videos, including YouTube Faces dataset [70], Point and Shoot Face Recognition Challenge (PaSC) dataset [74], and

Celebrity-1000 [75].

Although tracking techniques [76–78] can be used to associate the face images of a target in single-shot videos by utilizing the spatial, temporal, and appearance affinity, they are not effective for associating the target's face images present in multiple shots of a video. Hence, a video-based face identification technique that utilizes face tracking in a single shot cannot fully exploit useful information contained in multiple-shot videos, such as news videos, sport broadcasts [79], and movie trailers [80]. Figure 4.1 shows the target in a news video of multiple shots taken in several venues. An intra-shot face association technique should establish the linkage between face images within a video shot, and an inter-shot face association method should retrieve relevant face images from a single target annotation across multiple shots. Hence, the problem of performing video-based face identification task, where the target is only annotated once in the multiple-shot video, needs further investigation.

We propose a target face association (TFA) method to retrieve a set of representative face images in a video that have the same identity as the target. This set of associated face images is then utilized to generate a representation for face identification (See Figure 4.2). The TFA method leverages a linear support vector machine (SVM) to obtain the associated face images in the video. This linear SVM is trained iteratively with positive and negative instances guided by the cannot-link constraints. Note that several prior works have utilized the cannot-link constraints to learn effective metrics and models [20,23,81–84]. Initially, only face images corresponding to the target annotation are treated as the positive instances. The negative

Figure 4.1: A set of frames from a probe video in the JANUS CS3 dataset. This video consists of multiple shots taken from four scenes. The target is annotated with a red bounding box in frame #181, and faces extracted by the face detection algorithm are shown in green bounding boxes.

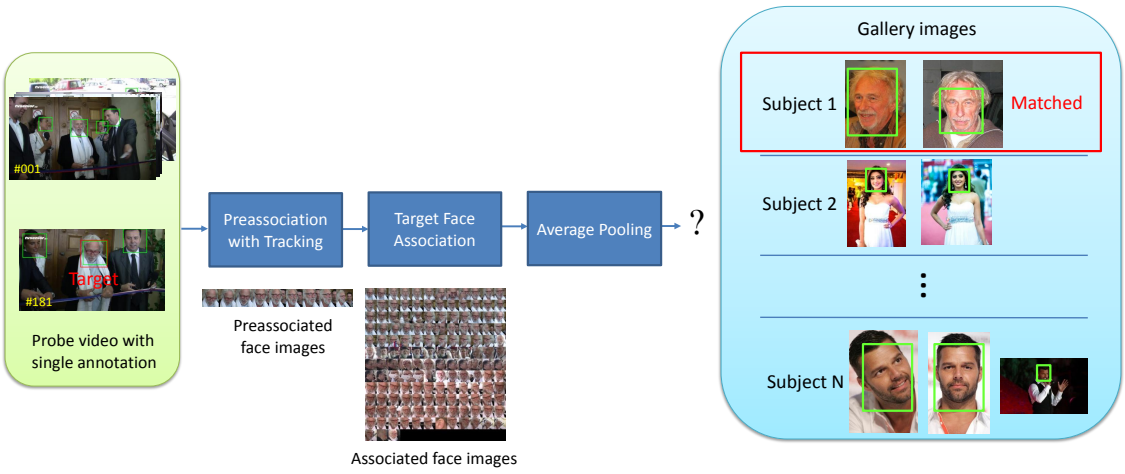


Figure 4.2: Video-based face association and identification.

training instances are the target's cannot-link instances, i.e., face images that appear with positive instances in the same frame. Face images that are classified as positive will undergo a pruning process to iteratively remove the least likely positive instance that violates the cannot-link constraint. Hence, the updated positive instances as well as their cannot-link negative instances can be used to update the linear SVM. If there is no negative training instance inferred by the target's cannot-link instances, we utilize an external face dataset as background negative instances. The idea of using a single instance against a large set of negative instances to learn the similarity function with respect to background subjects is an essential component in computing one-shot similarity [70, 85, 86]. Hence, we can learn a target-specific classifier by leveraging the background statistics in scenarios that do not have any within-video negative training instances.

To demonstrate the effectiveness of the proposed approach, we evaluate it with the recently released JANUS challenge set 3 (JANUS CS3) dataset, which is an extended version of [87]. In this dataset, a subject is represented by a template which contains images or videos from various media sources. A template is a succinct folder for organizing the exemplars of a subject in probe and gallery media. For instance, the FBI's wanted list usually has several images for a suspect [87]. In particular, the Protocol 6 of the JANUS CS3 dataset is a video-to-image face identification task in the open-set setting, which provides an end-to-end benchmark to evaluate the effectiveness of representing the probe template from the video of single annotation and to demonstrate the capability of searching for the mated (same-identity) template in the gallery. The evaluation results show that the proposed method achieves good

performance for the video-based face identification task.

The rest of this chapter is organized as follows. In Section 4.1, we review recent face retrieval and identification techniques. Section 4.2 describes the proposed target face association method. In Section 4.3, we demonstrate experimental results with the Protocol 6 of the JANUS CS3 dataset. Finally, Section 4.4 concludes this work with a brief summary.

4.1 Related Work

Various methods have been proposed in the literature for video face retrieval [67, 80, 88, 89]. An end-to-end video face retrieval system is proposed in [88], where several processing steps are utilized to handle variations due to pose, illuminations, and expressions (PIE). Sivic *et al.* [67] proposed to retrieve the subject of interest using a set of images that exhibits extensive variations of exemplars. The set of images is created from intra-shot matching, and the shot retrieval of a subject is obtained from computing the chi-square distance among the set of face images. Besides, several face retrieval techniques based on sparse representations have been proposed in [80, 89] to improve the robustness of face recognition algorithms. Recently, deep learning methods [72, 90–93] have shown significant improvement for face recognition over handcrafted features [94, 95] since the features learned by the deep convolutional neural network (DCNN) using large scale labeled face dataset are robust to PIE variations. Recently, a neural aggregated network [73] has been proposed to aggregate multiple face images in the video for generating a compact

representation for video-based face recognition. Moreover, video-adapted DCNN features [83,84] finetuned with automatically discovered cannot-link face tracks have shown improvements for clustering the face tracks in multiple-shot videos.

Prior works have shown improvements for video-based face recognition and retrieval tasks based on bounding boxes provided by human annotation or associated using a face tracker. The problem of selecting a set of representative face images of the target face from a multiple-shot video for creating a robust face representation has not been well studied in these works. Our work is motivated by the recent success of template adaptation technique [86]. The template adaptation technique is a form of transfer learning method [96] that employs a linear classifier to learn a template-specific similarity function. Such template-specific similarity function has shown improvements for face verification and identification tasks. We propose the TFA technique that learns a target-specific linear classifier for obtaining exemplars of the target face in the video. With a single annotation of the target, the linear classifier of TFA can automatically learn from training instances inferred by cannot-link constraints in the video.

4.2 The Proposed Method

In the video-based face association and identification task, we are given a single annotation of the target face in a probe video. The objective is to retrieve a set of representative face images of the target in the video, and then this set of face images is utilized to create a face representation of the target face for searching its

corresponding subject in the gallery. In the probe video, the target face is indicated by the human-annotated face bounding box $\mathbf{b}_0$ in frame f_0. There are m bounding boxes discovered by a face detector in a video. These face bounding boxes are denoted as $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m$, which are present in frames $f_1, f_2, \ldots, f_m$, respectively. The feature corresponding to the face image in bounding box $\mathbf{b}_i$ is denoted as $\mathbf{x}_i$. We aim to learn a target-specific SVM that can be used to classify a set of face images for creating a face representation of the target. We describe the details of each component of the proposed method as follows.

4.2.1 Face Preassociation with Tracking

Since learning the target-specific SVM requires the target annotation as the initial positive training instance, a low-quality target annotation, such as noisy, badly illuminated, and extreme pose face images, prevents the TFA from learning an effective SVM. A tracking technique is able to model the appearance and motion of a human head, and thus it allows us to capture subsequent face images of high quality for good initial representation. Hence, we employ an off-the-shelf tracking technique [97] to track the target face, and the face detection bounding boxes preassociated by tracking are utilized as the initial positive training set.

In each frame, the face detection bounding box that has the highest intersection-over-union (IoU) ratio [98] with the tracking bounding box is utilized as the preassociated face images of the target. As the tracking technique becomes vulnerable to severe occlusion and abrupt motion, we only incorporate those preassociated face

Figure 4.3: Preassociated face images using tracking. The first row shows the target annotation in videos, and the second row shows the preassociated face images using tracking.

detection bounding boxes in the first k frames. Since tracking across the shot boundary can lead to unexpected preassociation of face images from different subjects, we utilize a simple shot detection method by checking the absolute difference of pixel values between two consecutive frames. When the absolute difference is larger than a certain threshold, the preassociation with tracking is terminated. Figure 4.3 shows several cases where face preassociation with tracking improves the initial representation where the target annotation is corrupted due to extreme pose, noise, and occlusion. With the preassociation of tracking, we can obtain a set of preassociated face images as initial positive training instances to learn the linear SVM. The index set of the positive training images is $S_p = \{0\} \cup T$, where T consists of indices of those face images preassociated by tracking.

4.2.2 Target Face Association

We train a target-specific linear SVM from face images in the video to establish the intra/inter-shot face association of the target face. With the annotation of the target face and the preassociated face images, the index set of positive instances is initially represented as $S_p = \{0\} \cup T$. The negative training instances can be automatically discovered by utilizing the fact that the presence of a subject is unique. We define the cannot-link relation between the i^{th} and j^{th} face image as

$$g_{i,j} = \begin{cases} 1, & \text{if } r_{i,j} \leq \gamma,\ f_i = f_j,\ \text{and } i,j \in \{0,1,\ldots,m\}, \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

where $r_{i,j}$ is the IoU ratio between bounding box $\mathbf{b}_i$ and $\mathbf{b}_j$. Since the non-maximal suppression of the face detection response is not perfect, a face can be discovered by more than one bounding box. We set a tolerance threshold γ for the IoU ratio to avoid face images of similar bounding boxes in a frame being mistakenly enforced by the cannot-link constraints. Thus, $g_{i,j} = 1$ indicates that the i^{th} and j^{th} face image are far apart and appear in the same frame, and both images should not be identified as the same subject. Hence, the index set of within-video negative instances is represented as $S_n = \cup_{j \in S_p} \{i \mid g_{i,j} = 1\}$.

We introduce a background negative set $\{\mathbf{x}_i\}_{i=m+1}^{m+l}$ of l instances collected from an external face dataset to model the background subjects, and its corresponding index set is represented as $S_b = \{m+1, m+2, \ldots, m+l\}$. The background negative set becomes essential when there is no within-video training instance. We train a linear SVM with training data $\{(\mathbf{x}_i, y_i) | i \in (S_p \cup S_n \cup S_b)\}$, where the data label y_i

is expressed as

$$y_i = \begin{cases} 1, & \text{if } i \in S_p, \\ -1, & \text{otherwise.} \end{cases} \tag{4.2}$$

We propose two models to learn the weight vector $\mathbf{w}$ of the linear SVM.

4.2.2.1 Model 1

The weight vector $\mathbf{w}$ of the linear SVM is solved using the max-margin framework

$$\begin{aligned} \min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} &+ C_p \sum_{i \in S_p} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2 \\ &+ C_n \sum_{i \in S_n} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2 \\ &+ C_b \sum_{i \in S_b} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2, \end{aligned} \tag{4.3}$$

where

$$\begin{aligned} C_p &= C \frac{|S_p| + |S_n| + |S_b|}{2|S_p|}, \\ C_n &= C \frac{|S_p| + |S_n| + |S_b|}{2(|S_n| + |S_b|)} \frac{|S_n| + |S_b|}{2|S_n|} = \frac{|S_p| + |S_n| + |S_b|}{4|S_n|}, \text{ and} \\ C_b &= C \frac{|S_p| + |S_n| + |S_b|}{2(|S_n| + |S_b|)} \frac{|S_n| + |S_b|}{2|S_b|} = \frac{|S_p| + |S_n| + |S_b|}{4|S_b|} \end{aligned} \tag{4.4}$$

account for the weights to compensate for the class imbalance, and C is the cost parameter in the linear SVM. The weights are inversely proportional to the number of instances in positive and negative training sets. Among the negative samples, the weights are designed to be inversely proportional to the number of instances in S_n and S_b such that the importance of the within-video negative instances and background negative instances are balanced. We normalize $\mathbf{x}_i$ to unit norm, and then concatenate it with one to account for the bias. The normalized and augmented

feature vector is represented as

$$\bar{\mathbf{x}}_i = [\mathbf{x}_i^T / \|\mathbf{x}_i\| \quad 1]^T. \tag{4.5}$$

4.2.2.2 Model 2

Unlike Model 1 where the background negative instances are always utilized for training, we propose Model 2 that only utilizes the background negative instances when there is no within-video negative instance. The weight vector $\mathbf{w}$ of the linear SVM is solved using the max-margin framework

$$\begin{aligned} \min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_p \sum_{i \in S_p} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2 \\ + \mathbb{1}[S_n \neq \varnothing] C_n \sum_{i \in S_n} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2 \\ + \mathbb{1}[S_n = \varnothing] C_b \sum_{i \in S_b} \max[0, 1 - y_i \mathbf{w}^T \bar{\mathbf{x}}_i]^2, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} C_p &= \mathbb{1}[S_n \neq \varnothing] C \frac{|S_p| + |S_n|}{2|S_p|} + \mathbb{1}[S_n = \varnothing] C \frac{|S_p| + |S_b|}{2|S_p|}, \\ C_n &= C \frac{|S_p| + |S_n|}{2|S_n|}, \text{ and} \\ C_b &= C \frac{|S_p| + |S_b|}{2|S_b|}. \end{aligned} \tag{4.7}$$

Note that $\mathbb{1}[\cdot]$ is the indicator function. In this model, the negative training set is composed of the within-video negative instances that appear with positive instances in a frame. If there is no within-video negative instance, we employ the background negative instances as negative training instances.

The face images in the video that are classified as positive will be regarded as the associated face images of the target. In certain cases, the human-annotated

instance $\mathbf{x}_0$ can be misclassified as negative due to noise and extreme pose of a face image. Hence, we enforce the index 0 to be included in A, and the index set of the associated face images is represented as

$$A = \{0\} \cup \{i \mid \mathbf{w}^T \bar{\mathbf{x}}_i > 0, i = 1, \ldots, m\}. \tag{4.8}$$

The associated face images in A are assumed to have the same identity as the target. Nevertheless, face images in the same frame can be erroneously classified as the same subject and considered as the associated images in A. We propose to resolve such a conflict by iteratively removing the least likely instance among those instances that violate the cannot-link constraints.

The index set of those instances that violate the cannot-link constraints is represented as

$$Q = \{i \mid g_{i,j} = 1, i \in A, j \in A\}. \tag{4.9}$$

We can obtain the index of the least likely instance in Q by solving

$$\alpha = \underset{i \in Q - \{0\}}{\operatorname{argmin}} \ \mathbf{w}^T \bar{\mathbf{x}}_i. \tag{4.10}$$

To prevent the human-annotated instance $\mathbf{x}_0$ from being removed in A, we restrict the feasible space of (4.10) to $Q - \{0\}$. The index set of the associated face images can be updated by

$$A \leftarrow A - \{\alpha\}. \tag{4.11}$$

The above procedure is performed iteratively until all the violations are resolved.

4.2.3 Representation of the Target Face

Since the associated face images are assumed to have the same identity as the target does, we can use the associated face images as positive training instances ($S_p \leftarrow A$). With the cannot-link constraints, we can update the index set of the negative training instances by $S_n \leftarrow \cup_{j \in S_p} \{i \mid g_{i,j} = 1\}$. We can alternately update the associated face images in A and the weight vector $\mathbf{w}$ until the index set of the associated face images converges or the maximum number of iterations t_{max} is attained. The detailed procedure of the TFA is described in Algorithm 4. The associated face images in A may introduce outliers that do not have the same identity of the target face. To improve the reliability of the representation, we use the average pooling of the feature vectors of the associated face images to create the representation of the target face. We can express the representation of the target face as

$$\mathbf{x}^{fa} = \frac{1}{|A|} \sum_{i \in A} \mathbf{x}_i. \tag{4.12}$$

Note that the proposed method can handle the intra/inter-shot association of face images and face tracks. Although $\mathbf{x}_i$ represents a descriptor of a face image in this work, the proposed method can be easily extended to operate on track-level face descriptors [69, 71] and the cannot-link constraints among face tracks.

Algorithm 4 The algorithm for TFA

Input: $\{\mathbf{x}_i\}_{i=0}^{m+l}$, $\{\mathbf{b}_i\}_{i=0}^{m}$, $\{f_i\}_{i=0}^{m}$.

Initialization:

1: Establish the cannot-link constraints with (4.1);

2: Create the index set of preassociated face images T;

3: $S_p = \{0\} \cup T$, $S_n = \cup_{j \in S_p} \{i \mid g_{i,j} = 1\}$,

$S_b = \{m+1, m+2, \ldots, m+l\}, t = 0$;

4: **while** not converged and $t < t_{max}$ **do**

5: Obtain $\mathbf{w}$ with Model 1 using (4.3) or Model 2 using (4.6);

6: Update A with (4.8);

7: ▷ Line 8-12: Resolve the violations with cannot-link constraints

8: **while** not converged **do**

9: Update Q with (4.9);

10: Obtain α with (4.10);

11: $A \leftarrow A - \{\alpha\}$;

12: **end while**

13: $S_p \leftarrow A$;

14: $S_n \leftarrow \cup_{j \in S_p} \{i \mid g_{i,j} = 1\}$;

15: $t \leftarrow t + 1$;

16: **end while**

Output: A

4.3 Experimental Results

The JANUS CS3 dataset is the extended version of [87] which contains 11,876 images and 7,245 video clips of 1,870 subjects for evaluating face verification and identification task. We evaluate the proposed approach on Protocol 6 of the JANUS CS3 dataset. In Protocol 6, there are 7,195 probe templates. Each probe template provides a human-annotated face bounding box in a frame to mark the target in the video. These probe templates are evaluated with respect to two galleries. There are 940 and 930 templates in Gallery 1 and Gallery 2, respectively. The subjects in Gallery 1 and Gallery 2 are disjoint. Each template in the gallery consists of several images, and each image has a human-annotated bounding box provided to mark the subject in the image. The objective is to search the mated template in the gallery for a given probe template.

Protocol 6 is an open-set identification problem, and thus some of the probe templates will not have a mated template in the gallery. The ranking accuracy is evaluated with those probe templates that have a mated template in the gallery, which demonstrates the performance of closed-set search. To prevent the algorithm from using the prior knowledge that a probe template can always find its mated template in the gallery, the true positive identification rate (TPIR) and false positive identification rate (FPIR) are evaluated to demonstrate the performance of open-set search. The mathematical expressions of TPIR and FPIR can be found in [99]. Hence, a robust face identification technique should achieve high ranking accuracy as well as high TPIR at a specific FPIR.

For each probe template, we employ HyperFace [100] to discover the face bounding boxes in the video for every fifth frame as well as the target-annotated frame f_0, and the confidence threshold of the HyperFace detection is set at -0.5. Also, we refine the human-annotated bounding box of the target using the HyperFace detector. Face images are aligned with facial landmarks provided by HyperFace, and the aligned face images are represented by DCNN features [92, 93]. Each probe template with a single annotation in the video is converted to a representation of the target using TFA.

We use the kernelized correlation filter (KCF) tracker [97] to preassociate the face detection bounding boxes with the target face. The tracking algorithm is applied frame-by-frame to preassociate the face detection bounding boxes in every fifth frame for the first $k = 50$ frames. We notice that using more than 50 frames does not further improve the performance of video-based face identification performance. To prevent the unexpected preassociation resulting from the drifting of the tracker, a preassociated face detection bounding box whose IoU ratio with the tracking bounding box is less than 0.3 is discarded. For the parameters of TFA, two bounding boxes with IoU ratio less than $\gamma = 0.1$ are enforced by a cannot-link constraint, and the maximum number of iterations t_{max} is set at 3. We collect 160,498 face images from 1,710 subjects (1710-subject face dataset) to model the negative background subjects. Each subject is represented by the average of the feature vectors of a subject, and thus there are 1,710 instances in the background negative set. We use the weighted Liblinear implementation [101] with L2-regularized L2-loss support vector classification (primal) setting to learn the weight vector, and the cost parameter C

is set to 10.

The face representation of the i^{th} probe template $\mathbf{x}_i^{fa}$ is computed by (4.12), and the cosine similarity score between the i^{th} probe template and j^{th} gallery template is denoted as

$$s_{i,j} = cos(\mathbf{x}_i^{fa}, \mathbf{x}_j^{gal}), \tag{4.13}$$

where $\mathbf{x}_j^{gal}$ denotes the average of the feature vectors in the j^{th} gallery template. We compute two similarity matrices with two types of triplet embedded DCNN features [92, 93], respectively. The average of these two similarity matrices is used for performance evaluation. Note that the two types of DCNN features are trained with the CASIA-WebFace dataset [102]. The DCNN in [92] is trained with face images cropped with a tight bounding box, and the dimension of the feature is 320. On the other hand, the DCNN in [93] is trained with face images cropped with a loose bounding box that includes more context, and the dimension of the feature is 512. Both types of DCNN feature are embedded with the triplet probabilistic embedding matrix trained with the aforementioned 1710-subject face dataset, and the embedded feature dimension of each type of DCNN feature is 128.

We employ two baseline schemes for comparison. *Baseline 1* uses the target-annotated face image to represent the probe template. *Baseline 2* uses the face images preassociated by KCF tracker to represent the probe template. Note that the preassociation in *Baseline 2* is performed from the target-annotated frame to the upcoming shot boundary. Tables 4.1 and 4.2 present the face identification results on Gallery 1 and Gallery 2, respectively. The performance of probe videos

evaluated with Gallery 1 is substantially better than that with Gallery 2, but the image quality in Gallery 1 is not perceivably different from that of Gallery 2. One explanation is that the probe videos that have a mated template in Gallery 2 (Figure 4.4(b)) are more challenging to match than those in Gallery 1 (Figure 4.4(a)). This is evidenced by the fact that the target annotation in Figure 4.4(b) is of lower quality than that in Figure 4.4(a) in terms of resolution, illumination, pose, and occlusion. These unfavorable factors may lead to the performance degradation of TFA and face identification.

Table 4.3 presents the average results of two galleries. The performance of *Baseline 2* is better than that of *Baseline 1* since preassociation with KCF tracker is able to collect additional face images in the intra-shot for creating a diverse face representation. The proposed TFA method with Model 1 and Model 2 outperform the two baseline schemes since a face representation can be created from the associated face images in the intra-shots and inter-shots, which demonstrate more diverse representations than face images in the intra-shots alone. We observe that the performance of TFA (Model 2) is better than that of TFA (Model 1). The major difference between Model 2 and Model 1 is that the background negative instances are only utilized in Model 2 when there is no within-video negative instance. On the other hand, Model 1 always utilizes the background negative instances as part of negative training instances. This shows that the within-video negative instances are more effective than the background negative instances to train the decision boundary, and the target-specific classifier in Model 2 is more discriminative in separating the target from others by directly utilizing the within-video negative instances as

the negative training instances. Besides, background negative instances may have a different distribution as instances in the video, which introduces the unfavorable domain mismatch. Table 4.4 presents the identification results with various numbers of iterations using TFA. It is clear that most of the improvement is attained at the first iteration, and the improvement becomes marginal as the number of iterations increases.

We observe that the performance of TFA (Model 2) is 0.3% better on Rank-1 accuracy than TFA (Model 2, no preassociation). This shows that preassociation with KCF tracker is effective in improving the initial set of positive instances for learning a robust linear classifier. Although TFA is effective in retrieving most of the associated face images of the target to increase the diversity of the subject representation, it inevitably introduces some outliers and thus affects the identification accuracy. Hence, the advantage of diverse representations from the associated face images can be slightly offset by those incorrectly associated face images. Figure 4.5 shows the results of TFA in subsets of frames from three videos, and Figure 4.6 shows the associated face images corresponding to videos in Figure 4.5. Since the association quality of TFA highly depends on the performance of DCNN features and the characteristics of the face detector, we can observe that face images of extreme pose, blur, and illumination as well as false positives and false negatives of face detection have significant impact on the performance of TFA. Nevertheless, the proposed framework is general and can be adapted to work with any other face representation or face detector. Thus, it is expected to further improve the results when improved appearance features and face detectors are used.

Table 4.1: Results on Gallery 1 in the Protocol 6 of the JANUS CS3 dataset.

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Baseline 1:	0.5645	0.6963	0.7464	0.8094	0.8504	0.4654	0.2963
Baseline 2:	0.6323	0.7548	0.8010	0.8550	0.8904	0.5377	0.3725
TFA (Model 1, no preassociation)	0.6428	0.7593	0.7997	0.8515	0.8852	0.5496	0.3798
TFA (Model 2, no preassociation)	0.6650	0.7774	0.8195	0.8755	0.9106	0.5670	0.3760
TFA (Model 1)	0.6504	0.7704	0.8125	0.8623	0.8950	0.5576	0.3833
TFA (Model 2)	**0.6689**	**0.7875**	**0.8264**	**0.8803**	**0.9130**	**0.5701**	**0.3892**

Table 4.2: Results on Gallery 2 in the Protocol 6 of the JANUS CS3 dataset.

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Baseline 1:	0.4803	0.6014	0.6553	0.7218	0.7725	0.3773	0.2477
Baseline 2:	0.5183	0.6433	0.6917	0.7583	0.8079	0.3863	0.2630
TFA (Model 1, no preassociation)	0.5410	0.6641	0.7218	0.7755	0.8178	0.4208	**0.3032**
TFA (Model 2, no preassociation)	0.5493	0.6780	0.7292	0.7898	0.8373	0.4213	0.2926
TFA (Model 1)	0.5468	0.6745	0.7264	0.7810	0.8245	0.4204	0.3002
TFA (Model 2)	**0.5514**	**0.6803**	**0.7315**	**0.7926**	**0.8394**	**0.4245**	0.2931

Table 4.3: Average results of Gallery 1 and 2 in the Protocol 6 of the JANUS CS3 dataset.

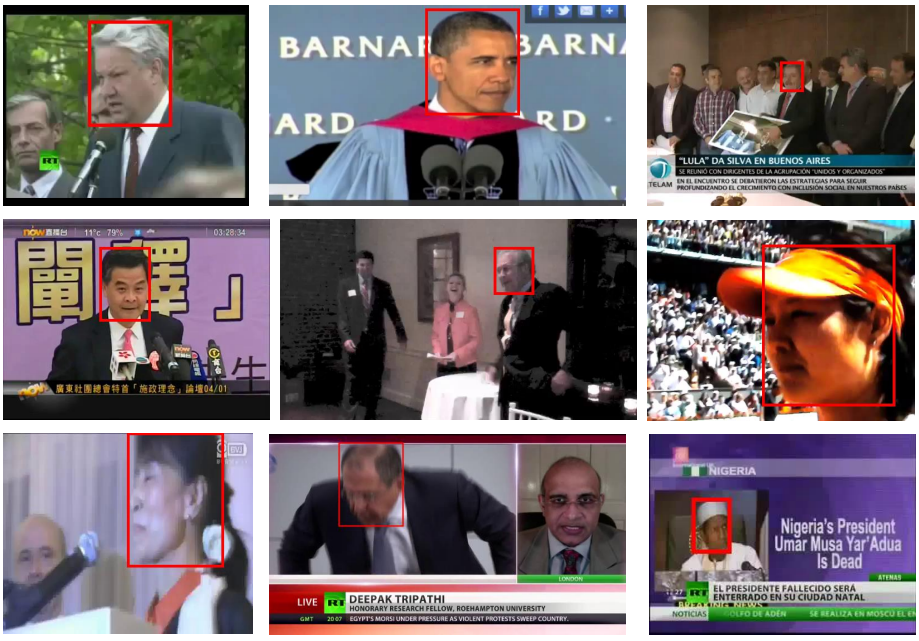
	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Baseline 1:	0.5224	0.6489	0.7009	0.7656	0.8114	0.4214	0.2720
Baseline 2:	0.5753	0.6990	0.7464	0.8066	0.8492	0.4620	0.3177
TFA (Model 1, no preassociation)	0.5919	0.7117	0.7607	0.8135	0.8515	0.4852	0.3415
TFA (Model 2, no preassociation)	0.6072	0.7277	0.7743	0.8326	0.8739	0.4941	0.3343
TFA (Model 1)	0.5986	0.7225	0.7695	0.8216	0.8597	0.4890	**0.3418**
TFA (Model 2)	**0.6101**	**0.7339**	**0.7790**	**0.8365**	**0.8762**	**0.4973**	0.3411

Table 4.4: Performance of TFA (Model 2) versus the number of iterations. We report the average results of Gallery 1 and 2 in the Protocol 6 of the JANUS CS3 dataset.

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Iteration 0 (i.e., $A = \{0\} \cup T$)	0.5594	0.6840	0.7289	0.7950	0.8371	0.4460	0.2873
Iteration 1	0.6080	0.7285	0.7750	0.8310	0.8695	0.4969	0.3389
Iteration 2	0.6102	0.7335	0.7791	0.8367	0.8743	0.4979	0.3391
Iteration 3	0.6101	0.7339	0.7790	0.8365	0.8762	0.4973	**0.3411**
Iteration 4	**0.6103**	0.7342	0.7789	0.8370	0.8767	**0.4981**	**0.3411**
Iteration 5	0.6100	**0.7346**	**0.7794**	**0.8377**	**0.8769**	0.4978	0.3393

4.4 Summary

In this chapter, we present the TFA approach to assist the video-based face identification task. With a single annotation of the target in the video, TFA can retrieve a set of representative face images in the video to create a representation of the target. Unlike tracking techniques that handle the association of face images in a video shot, the proposed method is capable of associating the face images across multiple shots in a video. The association is established by a target-specific linear classifier trained with face images of the target and background subjects in the video. The linear classifier is trained iteratively with the target's associated face images and the target's cannot-link face images. This target-specific linear classifier retrieves a set of face images to construct the representation of the target. Experimental results show that the target representation constructed by the associated face images is able to improve the performance of video-based face identification.



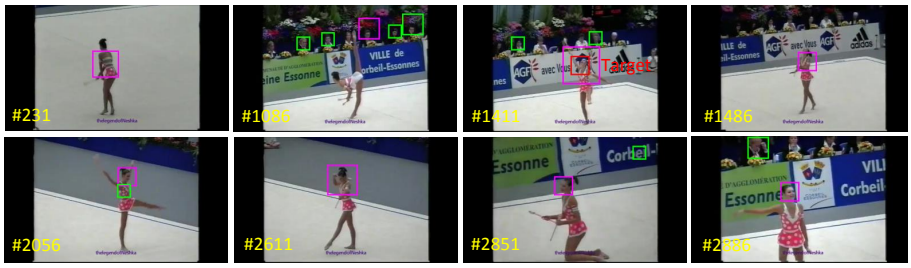
(a)

(b)

Figure 4.4: Target-annotated frames in the videos of JANUS CS3 dataset. (a) A subset of probe videos that has a mated template in Gallery 1. (b) A subset of probe videos that has a mated template in Gallery 2.

(a)

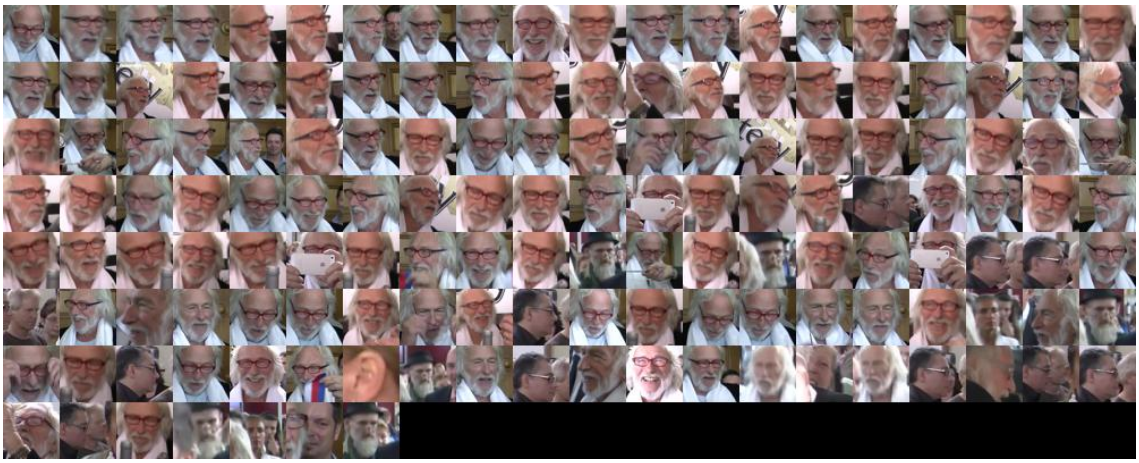
(b)



(c)

Figure 4.5: Subsets of frames that illustrate the associated face images of three videos in the JANUS CS3 dataset. The human-annotated bounding box of the target is shown in red, and the bounding boxes of the associated face images are shown in magenta.

(a)



(b)

(c)

Figure 4.6: Associated face images of three videos. Face images are displayed from top to bottom in the order of the confidence of face association.

Chapter 5: Face Recognition Using an Outdoor Camera Network

Outdoor camera networks have several applications in surveillance and scene understanding. Several prior works have investigated multiple person tracking [103–105], analysis of group behaviors [106, 107], anomaly detection [108], person re-identification [109], and face recognition [110–113] in camera networks. Face recognition in outdoor camera networks is particularly of interest in surveillance system for identifying persons of interest. Besides, the identities of subjects in the monitored area can be useful information for high-level understanding and description of scenes [114]. As persons in the monitored area may be non-cooperative, the face of a person is only visible to a subset of cameras. Hence, information collected from each camera should be jointly utilized to determine the identity of the subject. Unlike person re-identification, face recognition usually requires high-resolution images for extracting the detailed features of the face. As human faces possess a semi-rigid structure, this enables the face recognition method to develop 3D face models and multi-view descriptors for robust face representation.

Camera networks can be categorized into static camera networks and active camera networks. In static camera networks, cameras are placed around the monitored area with preset field of views (FOVs). The appearance of a face depends

on the relative viewpoints observed from the camera sensors and the potential occlusion in the scene, which has direct impact on the performance of recognition algorithms. Hence, prior work in [115] has proposed a method for optimal placement of static cameras in the scene based on the visibility of objects. Active vision techniques have shown improvements for the task of low-level image understanding than conventional passive vision techniques [116] by allocating resources based on current observations. Active camera networks usually comprise of a mixture of static cameras and pan-tilt-zoom (PTZ) cameras. During operation, PTZ cameras are continuously reconfigured such that the coverage, resolution (target coverage), informative view, and the risk of missing the target are properly managed to maximize the utility of the application [111, 112].

A recent research survey on active camera network is provided in [117], and the authors propose a high-level framework for dynamic reconfiguration of camera networks. This framework consists of local cameras, fusion unit, and a reconfiguration unit. The local cameras capture information in the environment and submit all the information to the fusion unit. The fusion unit abstracts the manipulation of information from local cameras in a centralized or distributed processing framework and outputs the fused information. The reconfiguration unit optimizes the reconfiguration parameters based on fused information, resource constraints, and objectives. In centralized processing frameworks, the information from each camera is conveyed to a central node for predicting the states of the observations and reconfiguring the local cameras. On the other hand, the distributed processing of the camera networks becomes desirable when the bandwidth and power resources are limited. In

this scenario, each camera node receives information from its neighboring nodes and performs the tasks of prediction and reconfiguration locally.

Face association across video frames is an important component in any face recognition algorithm that processes videos. When there are multiple faces appearing in a camera view, robust face-to-face association methods should track the multiple faces across the frames and avoid the potential of identity switching. Also, face images observed from multiple views should be properly associated for effective face recognition. When the cameras are calibrated, the correspondence of face images observed in multiple views can be established by geometric localization methods, e.g., triangulation. Nevertheless, geometric localization methods demand accurate calibration and synchronization among the cameras, and they usually require the target to be observed by at least two calibrated cameras. Hence, these methods are not suitable for associating face images captured by a single PTZ camera operating at various zoom settings. Alternatively, the association between face images observed in multiple views can be established by utilizing the appearance of upper body [23, 113, 118]. This method is effective as the human body is more perceivable than the face. Besides, the visibility of human body is not restricted to certain view angle as the human face does. Based on this fact, a face-to-person technique has been developed in [113] to associate the face in the zoomed-in mode with the person in the zoomed-out mode. In order to effectively utilize all the captured face images for robust recognition, face-to-face and face-to-person associations have become the fundamental modules to ensure that the face images captured from different cameras and various FOVs are correctly associated.

Face images captured by cameras in outdoor environments are often not as constrained as mug shots since persons in the scene are typically non-cooperative. Furthermore, the face images captured by cameras deployed in outdoor environments can be affected by illumination changes, pose variations, dynamic backgrounds, and occlusions. Moreover, the sudden changes in PTZ settings in active camera networks can introduce motion blur. Although constructing a 3D face model from face images enables synthesis of different views for pose-invariant recognition, it typically relies on accurate camera calibration, synchronization, and high-resolution images. Hence, we address several issues that come up while designing a face recognition algorithm for outdoor camera systems. The objective is to extract diverse and compact face representation from multi-view videos for robust recognition. Also, context information, such as gaze, activity, clothing appearance, and unique presence, can provide additional cues for improving the recognition performance.

In this chapter, we first review the taxonomy of camera networks in Section 5.1. Techniques for face association are discussed in Section 5.2. Several issues for face recognition using images and videos captured in outdoor environments are discussed in Section 5.3. In Section 5.4, we present the design details of a camera network system for face recognition. Some remaining challenges in outdoor camera networks are presented in Section 5.5. We conclude the chapter in Section 5.6.

5.1 Taxonomy of Camera Networks

Several designs of camera networks have been developed to facilitate multiple camera-based surveillance systems. Camera networks can be categorized into **static camera networks** and **active camera networks**. Characteristics of camera networks, such as the centralized/distributed processing framework and overlapping/non-overlapping camera network, will be discussed in this section.

5.1.1 Static Camera Networks

Static camera networks typically consist of multiple cameras mounted in fixed locations, and the preset FOVs of the cameras are not reconfigurable during operation. Static camera networks have been used in multiple person tracking [103–105] and person re-identification [109]. In order to enhance the coverage area, an omni-directional camera has been utilized along with a regular perspective camera [119]. There are very few works utilizing the static camera networks for remote face recognition in outdoor environments since a static camera lacks the zooming capability to capture the close-up view of faces. Some of the designs preset the static camera to the known walking path of pedestrians for capturing the facial details [120]. In practice, static camera networks for face recognition require densely distributed cameras to opportunistically capture the face images in a wide area. Prior work reported in [115] has proposed a strategy for the optimal placement of cameras to ensure that a face of interest is visible to at least two cameras. The objective is to maximize the visibility function among all the camera setup parameters (loca-

tions and FOVs) in consideration of potential occlusions in the scene. As the static camera often lacks the zooming capability to capture the close-up view, face images captured from a remote camera may not have sufficient resolution and good quality. Hence, remote face recognition [121] becomes one of the important issues in static camera networks.

5.1.2 Active Camera Networks

In active camera networks, cameras are reconfigurable during operation to maximize the utility of a certain application. Most of the active camera networks utilize a hybrid of static cameras and PTZ cameras, and the utility function can be formulated as the coverage for the face of interest or the appearance quality of faces [112]. A common setup in active camera networks is the master and slave configuration. Static cameras observe the wide area for performing the task of detection and localization. The PTZ cameras possess the flexibility to capture close-up views of faces. The master and slave camera networks usually adopt a centralized processing framework to reconfigure the slave cameras based on observations from the master camera. Active distributed PTZ camera networks have been proposed to collaboratively and opportunistically capture informative views and satisfy the coverage constraints [111, 112].

5.1.3 Characteristics of Camera Networks

The information collected by multiple cameras can be processed in a centralized or distributed framework. In the centralized processing framework, information from all the camera sensors is conveyed to a base station to estimate the tracking states and determine the identities. The distributed framework can reduce the amount of data transfer by processing the information locally and then convey the succinct information to other nodes. Given the limited resources of bandwidth and power in distributed camera networks, exchanging visual data among sensor nodes is not preferred. Hence, each sensor node only conveys modest information extracted from visual content to other sensor nodes. Based on the received information and its own visual content, each sensor node computes local optimal settings, e.g., PTZ settings of camera, to achieve the common goal.

For distributed camera networks in a wide area, cameras do not always have overlapping FOVs. Hence, the camera topology (connectivity between non-overlapping FOVs of cameras) should be established by exploring the statistical dependency on the entry and exit activities between cameras [122, 123]. Besides, spatial-temporal constraints and relative appearance of the persons can be utilized for persistent tracking in non-overlapping FOVs [124,125]. With the topology of a non-overlapping camera network, faces and persons appearing from one view to another can be successfully associated for robust recognition.

5.2 Face Association in Camera Networks

Face association relies on persistent person tracking and face acquisition in outdoor camera networks. In this section, we investigate face-to-face and face-to-person associations [113], which enable robust recognition in long-term and wide-area monitoring scenarios.

5.2.1 Face-to-face Association

A successful face-to-face association algorithm can continuously track the movement and appearance changes of faces over time. Nevertheless, face-to-face association is challenging since multiple faces appearing in the scene can introduce ambiguities. Especially, faces of a group of people are likely to be occluded by each other when the face images are captured from a single viewpoint. Hence, it is essential to correctly associate the face images to form face tracks, and then recognition can be performed effectively for each track based on the assumption that a face track only consists of face images captured from the same subject.

In general, a multiple face tracking algorithm handles the initialization of face tracks, simultaneous tracking of multiple faces, and the termination of a face track. There are several challenges to be addressed while designing a multiple face tracking algorithm. Face tracks that are spatially close to each other can lead to identity switching. The drift of face tracks can result due to large pose variations of faces. Besides, face tracks become fragmented due to occlusion and unreliable face detection. Moreover, videos captured by the hand-held cameras can be affected

by unexpected camera motion, which makes the association of face images difficult. Given the recent advancements in multiple object tracking (MOT) [126–128], several methods have utilized the framework of MOT for multiple face tracking [13, 77].

Roth *et al.* [77] adapted the framework of multiple object tracking to multiple face tracking based on tracklet linking, and several face-specific metrics and constraints were used for enhancing tracking reliability. Wu *et al.* [13] modeled the face clustering and tracklet linking steps using a Markov Random Field (MRF), and the fragmented face tracks resulting from occlusion or unreliable face detection were then associated to produce reliable face tracks. Duffner and Odobez [127] proposed a multi-face Markov Chain Monte Carlo (MCMC) particle filter and a Hidden Markov Model (HMM)-based method for track management. The track management strategy includes the creation and termination of tracklets. A recent work in [78] proposed to manage the track from the continuous face detection output without relying on long-term observations. In unconstrained scenarios, the camera can be affected by abrupt movements, which makes consistent tracking of faces challenging. Du and Chellappa proposed a conditional random field (CRF) framework to associate faces in two consecutive frames by utilizing the affinity of facial features, location, motion, and clothing appearance [23, 118].

Although linking of tracklets from the bounding boxes provided in face detection has shown some robustness in multiple face tracking, performing face detection for every frame is not feasible due to high demands on computational resources. As shown in Figure 5.1, the face association method in [129] detects faces for every 5 frames and uses the Kanade-Lucas-Tomasi (KLT) feature tracker [130] for short-

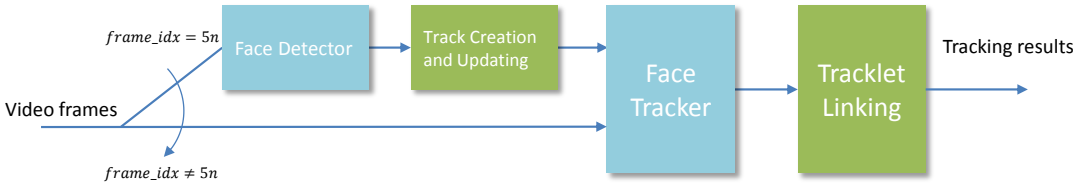


Figure 5.1: Block diagram of the multiple face tracking framework.

term tracking. The bounding boxes provided by detection and KLT tracking serve as inputs for the tracklet linking algorithm [131].

5.2.2 Face-to-person Association

Face recognition in camera networks requires the persistently tracked person and correct association of captured faces. In overlapping camera networks, the correspondence of faces captured from multiple views can be established from geometric localization methods, i.e., triangulation. Nevertheless, these techniques may not be applicable for non-overlapping camera networks.

For PTZ cameras in a distributed camera network, each PTZ camera actively performs face acquisition operating at different FOVs. It is essential to perform face-to-person association since the number of faces and the number of persons in the field of view may not be consistent when switching between zoomed-out and zoomed-in mode. Face-to-person association ensures that face images of a target captured from various FOVs can be registered with the same person for identification. The appearance of face images captured by the zoomed-in mode can be quite different

from that of full-body images captured by the zoomed-out mode since the close-up views only capture a portion of the full-body images. Hence, the HSV color histogram is used to model the appearance of upper-body in different zoom ratio [113], and the Hungarian algorithm [15] is employed to find the optimal assignment between faces and persons.

5.3 Face Recognition in Outdoor Environments

In this section, we discuss several issues when performing the recognition task on images and videos captured by outdoor camera networks.

5.3.1 Robust Descriptors for Face Recognition

Several techniques have been proposed to overcome many challenges due to pose variations by extracting handcrafted features around the local landmarks of face images, and a discriminative distance metric is learned such that a pair of face images from the same person will induce a smaller distance than that from different persons. Chen *et al.* [132] used multi-scale and densely sampled local binary pattern (LBP) features and trained the joint Bayesian distance metric [133]. Simonyan *et al.* performed Fisher Vector (FV) encoding on densely sampled SIFT feature [95] to select highly representative features. Li *et al.* [134] proposed a pose-robust verification technique by utilizing the probabilistic elastic part (PEP) model, and thus the impact of pose variations could be alleviated by establishing the correspondence between local appearance features (e.g. SIFT, LBP, etc.) of the two

face images. Recently, Li and Hua [135] proposed a hierarchical PEP model to exploit the fine-grained structure of face images, which outperforms their original PEP model. Recently developed methods based on handcrafted features and deep features [90, 91] have shown significantly improved performance. However, learning the deep features usually requires a large number of labeled training data.

As face images captured in outdoor environments suffer from low-resolution and occlusion, face alignment becomes challenging. Liao *et al.* [136] have proposed an alignment-free face recognition using multi-keypoint descriptors, and the size of the descriptor can adapt to the actual content.

5.3.2 Video-based Face Recognition

In camera networks, sequences of face images in videos can capture diverse views and facial variations of an individual (assuming a face track only consists of face images from one person). Hence, several works have proposed video-based methods for effective representations. Zhou *et al.* [137, 138] proposed to simultaneously characterize the appearance, kinematics and identity of human face using particle filters. Lee *et al.* [68] constructed the pose manifold from k-means clustering of face tracks and established the connectivity across the pose manifold for representing the face images in the video. Chen *et al.* [139] proposed to cluster a face track into several partitions, and dictionaries learned from each partition are used for capturing the pose variations of a subject. Li *et al.* [69] adopted the PEP model for constructing the video-level representation of face images. The video-level represen-

tation was computed by performing the pixel-level-mean of the PEP-representation of video frames. Most video-based verification techniques have extracted diverse and compact frame-level information for constructing video-level representation.

5.3.3 Multi-view and 3D Face Recognition

Robust face recognition depends on the availability of effective descriptions of faces. Several prior works have investigated the affine invariant features that are robust to slight pose variations or view changes. Nevertheless, the 2D model fails to represent large pose variations due to self-occlusion and the perspective distortion introduced when the face is close to cameras. The 3D model overcomes these disadvantages of the 2D model by describing the features on the 3D structure. Given the estimated pose of the face, the features of a face collected from multiple views can be jointly registered onto a 3D structure, and thus they are no longer dependent on the pose variation of the face itself. In face tracking or recognition, the variation of the head structure is modest. Hence, the 3D feature can be densely constructed by mapping facial textures onto a generic 3D structure. Nevertheless, successful modeling of 3D faces requires reliable camera calibration for accurate registration, which is generally not sufficiently precise in real world surveillance scenarios. In the following, we review prior works that exploit the multiple views and 3D face models for recognition.

An *et al.* [1] adopted the dynamic Bayesian network (DBN) for face recognition in camera surveillance network by encoding the temporal information and features

from multiple views. The DBN consists of a root node and several camera nodes in a time slice. Figure 5.2 shows the DBN structure using three cameras of three time slices. The root nodes capture the distribution of the subjects in the gallery, and the camera nodes contain the features of a face observed from each camera. The time slices enable the DBN to encode the temporal variations of a face. Du *et*

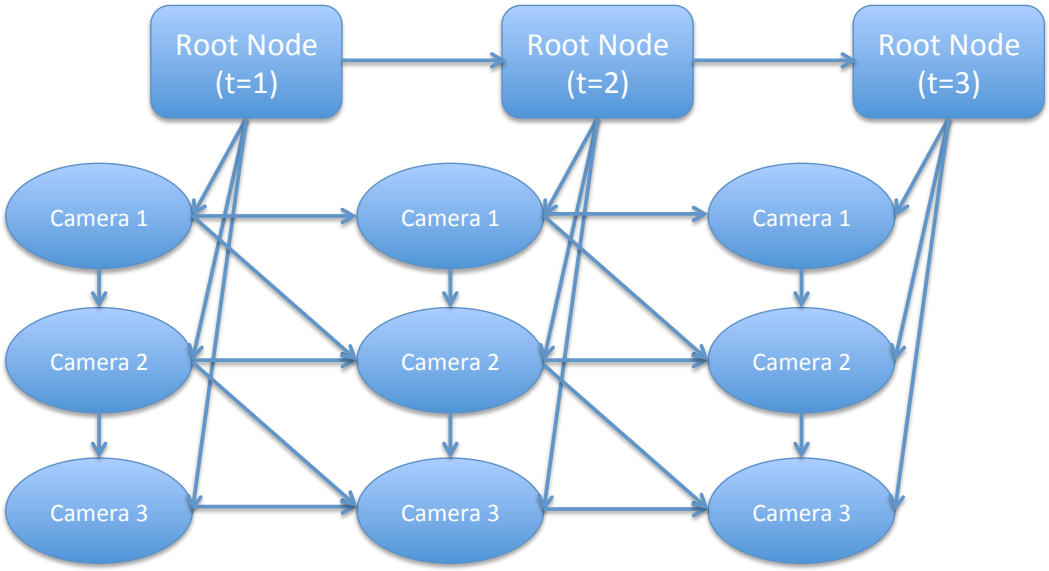


Figure 5.2: The DBN structure using three cameras of three time slices [1].

al. [2] proposed a robust face recognition method based on the spherical harmonic (SH) representation for the texture-mapped multi-view face images on a 3D sphere. Figure 5.3 shows the texture mapping on a 3D sphere from three cameras. The textured-mapped 3D sphere was used for computing the SH representation. The method is pose-invariant since the spectrum of the SH coefficients is invariant to the rotation of head pose.

Besides, several prior works have utilized structure-from-motion techniques to reconstruct the 3D model for face recognition from multiple face images [120, 140].

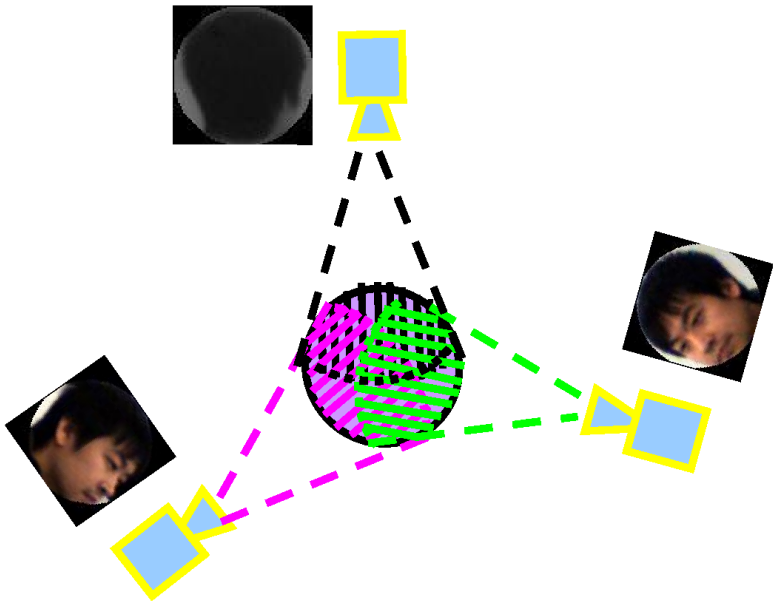


Figure 5.3: The spherical 2D face images captured from three cameras are mapped on to the 3D facial sphere, which will be used to compute the SH representation [2].

5.3.4 Face Recognition with Context Information

Context features, such as clothing, activity, attributes, and gait, can serve as additional cues for improving the performance of face recognition algorithms [141]. Moreover, the uniqueness constraint can be utilized to improve the recognition accuracy since two persons presenting in a venue should not be identified as the same subject. Liu and Sarkar [142] proposed a recognition framework by fusing the gait and face information, and several fusion strategies for integrating these two biometric

modalities were evaluated. Their experimental results show that the combination of one face and one gait per person gives better result than two face probes per person and two gait probes per person. This shows that different biometric modalities can be fused to further improve the recognition accuracy.

5.3.5 Incremental Learning of Face Recognition

Besides, the outdoor environment can change due to time of day, weather, etc., and thus the distribution of data can change. As a result, the model should adapt to the current captured data for effective face recognition. A recent work in [143] has proposed an adaptive ensemble method to alleviate the impact of environmental changes on face recognition by utilizing diversified learned models. The method first performed change detection to distinguish if the current input significantly differs from the learned model. Otherwise, a corresponding model is selected for recognition. Long-term memory was then used to store the parameters for identifying new concepts and training new model-specific classifier of each subject. The short-term memory holds the validation data for referencing. The system model can be updated by adopting the boosting-based method for learning independent classifiers and performing weighted fusion.

As the outdoor scene is an open environment, it is common that a subject does not belong to any subject in the gallery. Several works have addressed the issue of open set recognition [144, 145]. Subjects that have not been seen in the gallery should be rejected, and the captured face images can be potentially used for

learning models of new subjects.

5.4 Outdoor Camera Systems

In this section, we review several camera networks deployed in outdoor environments.

5.4.1 Static Camera Approach

Medioni *et al.* [120] used two static cameras to monitor a chosen region of interest. In this work, one of the static cameras provided high-resolution face images with a narrow FOV, and the other camera captured the full body of pedestrians in the scene with a wide FOV. A 3D face model was constructed from multi-view stereo technique operating on the sequences of face images. Stereo pair of wide baseline can be challenging for establishing correspondence but often provide better disparity resolution. On the other hand, it is easy to establish correspondence for a short baseline stereo pair, but the disparity resolution might be insufficient. The task involved key frame selection to form multiple stereo pairs from near frontal images within -10 to 10 degrees. Each pairwise stereo pair contributed to a disparity map that represents the height of the 3D face surface. The mesh descriptor of the 3D face model was obtained from integrating the multiple disparity maps, and outliers of disparity were rejected by surface self-consistency. The 3D face models and 2D face images were used for biometric recognition. Although this approach is capable of reconstructing the 3D face from outdoor video sequences, their experimental results

show that the performance of 3D face recognition degrades as the resolution of face images is reduced due to the increase in distance. This reveals that face recognition based on 3D modeling in outdoor environments remains a challenging task.

5.4.2 Single PTZ Camera Approach

Face recognition systems using a single PTZ camera are challenging to design since the persistent tracking of a person, camera control to follow the identity, and recognizing the identity from face images should be performed simultaneously. Dinh *et al.* [146] proposed a single PTZ camera acquisition strategy for extracting high-resolution face sequences of a single person. Their method employs a pedestrian detector in the wide FOV to detect face of interest. Once a pedestrian is detected, the pan-tilt parameters of camera are adjusted to bring the face of a pedestrian to the center of the image and the zoom parameter is preset to ensure sufficient resolution of the face images. As the face detector localizes a face, the active tracking mode is initiated by performing face tracking with the bounding box provided by the face detector. In the meantime, camera control is initiated to follow the face simultaneously. Since the tracked face consistently moves in the scene, the camera control module in [147] is employed to follow the target precisely and smoothly by sending commands to reconfigure the pan-tilt parameters. The zoom parameter is dynamically adjusted to ensure that a face in the FOV has a proper size. When the face drifts out of the FOV of a camera, the one-step-back strategy camera control is adopted by decreasing the focal length for one-step until the face reappears in the

FOV.

Cai *et al.* [113] employed a single PTZ camera for face acquisition for multiple persons in the scene. The PTZ camera switches between zoomed-in and zoomed-out mode for obtaining narrow and wide FOV, respectively. In the zoomed-out mode, person-to-person association was employed to track multiple persons in the scene. When the camera is switched from zoomed-out mode to zoomed-in mode to obtain the close-up view of a particular person, the face-to-person association was performed to ensure that the detected faces in the zoomed-in mode are correctly associated with the person in the zoomed-out mode. The face-to-face and face-to-person associations are employed when switching between zoomed-in and zoomed-out modes. The camera scheduling is based on a weighted round-robin mode to acquire close-up views of each person in the scene.

5.4.3 Master and Slave Camera Approach

In the University of Maryland (UMD), an outdoor camera network comprising of four sets of the off-the-shelf PTZ network IP cameras (Sony SNC-RH164) is employed to acquire face images in the open area in front of a campus building. A master and slave camera framework is adopted in the outdoor camera system. The PTZ cameras are deployed on the roof and side walls of the building as shown in Figure 5.4, and their corresponding locations seen from the top view are marked in the world map of Figure 5.5. One of the PTZ camera serves as the master camera (M) and other three cameras serve as slave cameras (S1, S2, and S3). The resolution

Figure 5.4: The deployment of PTZ cameras at the University of Maryland campus.

of the video stream from each camera is 640×368 pixels, and the frame rate is 15 frames/second.

The proposed system consists of several modules, including foreground detection [148], blob tracking, face detection, face recognition, and the surveillance interface.

5.4.3.1 Camera Calibration

Using the steerable functionality of PTZ camera, we calibrate the intrinsic parameters of each camera using techniques presented in [149] without using a known pattern [150]. During calibration, we steer all the PTZ cameras to look at a common

Figure 5.5: The interface of UMD outdoor camera network. The first column shows the view from the master camera, the world map, and the eight subjects in the gallery. The second column shows views from three slave cameras. The pedestrians in the view of master camera are tracked with bounding boxes, and their locations are marked on the world map. The predicted identity of each tracked pedestrian is annotated in the world map.

overlapping area, and all the cameras are zoomed out to maximize the overlapping FOVs. Since the perspectives are quite different across the different PTZ cameras, we manually select the common corresponding points (Figure 5.6) for extrinsic calibration [151]. The extrinsic parameters of the PTZ cameras are computed by the Bundler toolkit [152] to obtain the rotation and translation matrices relative to the master camera. Moreover, we assume that the pedestrian movement can be modeled as a planer motion on the ground plane, and thus we simply mark the location of the pedestrian in the world map. By using at least three (manually selected) 3D coordinates on the ground, we obtain the planar equation of the ground plane

$$aX + bY + cZ = d, \tag{5.1}$$

and the unit normal vector of the ground plane is denoted as $\mathbf{v}_n =< a, b, c > /\sqrt{a^2 + b^2 + c^2}$.

5.4.3.2 Camera Control

The objective of the outdoor camera system is to recognize the identity of a pedestrian in the area being monitored from a set of subjects in the gallery, and we report its location and identity in the world map as shown in Figure 5.5. In the view of the master camera, the moving pedestrians are first detected by the foreground detection, and then tracked by the blob trackers. We use the foreground detection and blob tracking methods provided in OpenCV [153]. The image coordinate at the standpoint of a pedestrian (x, y) is converted into the 3D world coordinate $\mathbf{x}_f$ to indicate the 3D coordinate of the foot of the pedestrian. The 3D world coordinate $\mathbf{x}_f$

(a) Master camera (M)

(b) Slave camera 1 (S1)

(c) Slave camera 2 (S2)

(d) Slave camera 3 (S3)

Figure 5.6: The common corresponding points (green crosses) in master and slave camera views are used for extrinsic calibration.

is computed by intersecting the ray along the homogeneous coordinates (xz, yz, z) of the master view with the planner equation of the ground in (5.1). In order to capture the high-resolution face images for recognition, the slave cameras are steered to point at the 3D coordinate of the head such that the head of the pedestrian are brought to the image center. We compute the rough 3D coordinate of the head $\mathbf{x}_h$ in the world as

$$\mathbf{x}_h = \mathbf{x}_f + h\mathbf{v}_n, \tag{5.2}$$

where h is the average human height of a pedestrian in the scene. In the system, it is empirically set as a constant. However, a more precise height of a pedestrian in the

scene can be computed from a reference object of known height and the vanishing point [154].

A simple camera scheduling strategy is implemented to steer all the slave cameras to point at the head of a person simultaneously. When there is more than one person in the monitored area, each person is sequentially observed by all the slave cameras with a time interval of 4 seconds. Sophisticated camera scheduling algorithm, such as [112], can be implemented to allocate the PTZ cameras to optimally capture the most informative views. Hence, PTZ cameras can be individually steered to capture the face images from different persons in parallel.

5.4.3.3 Face Recognition

The sequence of face images detected in the camera views are recognized by the video-based face recognition method developed by Chen *et al.* [139]. The dictionaries for the 8 subjects in the gallery are trained offline from two sessions of videos captured from three slave cameras. In the training stage, each face image in grayscale is resized to 30 $\times$ 30 pixels, and each face image is then vectorized into feature vector of dimension 900. Feature vectors of each subject are clustered into ten partitions using k-means clustering. Figure 5.7 shows a subset of partitions from three subjects in the gallery. There are ten sub-dictionaries of each subject learned from the ten partitions to build the compact face representation of each subject. In the testing stage, the identity of a face image in each frame is predicted by assigning the identity of the sub-dictionary that yields the minimum reconstruction error of

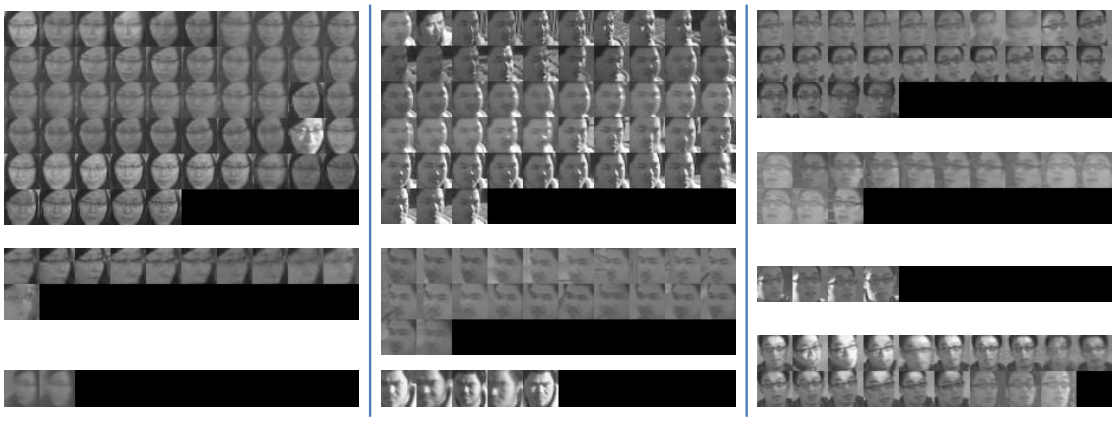


Figure 5.7: A subset of partitions from three subjects used for dictionary learning.

the face image. The identity of the pedestrian is then determined by using a majority voting that accounts for the predicted identity of face images of previous frames. The location and identity of each pedestrian is continuously updated on the world map.

5.4.4 Distributed Active Camera Networks

In master and slave camera networks, the functionality of each camera is assigned throughout the operation. On the other hand, in the collaborative and opportunistic PTZ camera networks, the tasks of tracking in wide FOV and capturing high-resolution images in narrow FOV are dynamically reconfigured based on the current observations. Each PTZ camera is capable of low-level processing, including target tracking and common consensus state estimation.

Ding *et al.* [112] have implemented distributed active camera networks of 5 and 9 PTZ cameras, which provide the dynamic coverage of the monitored area.

The configuration of the PTZ settings relies on a distributed tracking method based on the Kalman-Consensus filter [104, 105]. Neighboring cameras can communicate with each other and negotiate with neighboring nodes before taking an action. The framework optimizes the distributed camera configurations by maximizing the utility based on the specified tracking accuracy, informative shot, and image quality, in the active distributive camera network. The utility function can model the area coverage and target coverage. Another framework in [111] uses a camera network of 215 PTZ cameras to opportunistically retrieve informative views. A Bayesian framework is utilized to perform the trade-off between the reward of informative view and the cost of missing a target. Besides, a framework proposed by Morye *et al.* [155] continuously changes the camera parameters to satisfy the tracking constraint and opportunistically capturing the high-resolution faces. Image quality is formulated as a function of the target resolution and its relative pose with respect to the view camera.

5.5 Remaining Challenges and Emerging Techniques

Video surveillance in complex scenarios remains a challenging task since existing computer vision algorithms cannot adequately address the challenges due to pose variations, severe occlusions, illumination changes, and ambiguity between identities of similar appearance. Although the 3D structure of face can provide distinct features, the issues of synchronization error, calibration error, insufficient imaging resolution of remote identity, make it difficult to recover the 3D face model. The

challenges of designing a video surveillance systems do not depend on a single factor, and the performance of one stage can potentially suffer from unreliable results in previous stages. All these factors make face recognition in outdoor camera networks a challenging task.

Face recognition in mobile camera network is an emerging research topic [156]. In this scenario, each visual sensor is mounted on a mobile agent and works cooperatively with other visual sensors in the mobile networks. Given the limited bandwidth and power in the mobile networks, exchanging visual data between sensors becomes infeasible. Hence, each sensor node only conveys a modest amount of information extracted from a particular camera to other sensor nodes. Based on the received information and its own visual data, each sensor node computes an optimal setting such as the moving direction of the mobile agent or the PTZ setting of camera to achieve the common goal in the networks. With the low cost of drones, cameras mounted on flying mobile agents have been utilized for face recognition [157]. As compared to conventional mobile agents, drones are less restricted by the geographic constraints. Nevertheless, sophisticated drone stabilization techniques, camera controls, and communication techniques should be developed for conveying informative and stable face images for face recognition in drone-based video surveillance.

With the prevailing use of personal mobile devices, the utilization of camera sensors embedding GPS and orientation sensor remains an open problem to solve. Unlike typical mobile networks where the algorithm gets full control on the steering of mobile agent, the visual sensors on personal device usually acquire visual data passively. Hence, crowd-based services as part of the mobile camera network should

take into account the behavior of user and human interaction to opportunistically collect information for face recognition in large-scale and unrestricted environments.

5.6 Summary

In this chapter, we first discussed the usefulness of camera networks for face recognition in outdoor environments. The static camera networks are suitable for densely distributed wide area, but they are not as flexible as the active camera networks. The active camera networks can take advantages of the PTZ capability to opportunistically capture high-resolution face images. Nevertheless, face images captured in outdoor environments are unconstrained, and the quality is usually affected by pose variations, illumination changes, occlusions, and motion blur. Effective multi-view video-based methods should be employed to build diverse and compact face representations. We reviewed several issues relevant to the design of camera network systems for face recognition deployed in outdoor environments. Remaining challenges such as handling real-time operation, synchronization, etc., should be overcome to make the outdoor camera network systems for face recognition pervasive and reliable. Finally, we discussed the design details of a camera network-based system implements at the University of Maryland.

Chapter 6: Conclusions and Directions for Future Work

In this dissertation, we discussed the face recognition problem in scenarios where the training or testing data is weakly labeled. We proposed several robust techniques to overcome such imperfections by exploiting the non-local cost aggregation technique to reduce the impact of noisy labeling and utilizing the low-rank matrix approximation method to recover the actual labels. Besides, the target face association is capable of generating a set of associated face images of the target from a single human-annotated bounding box to support robust face identification. Several directions for extending are briefly summarized below.

6.1 Character identification in TV-series

We proposed a unified framework for character identification in a TV-series. We constructed the track nodes from face and person tracks, and the track nodes served as the basic elements in constructing the MST. As the track nodes with similar appearance become adjacent in the MST, we showed how the non-local cost aggregation method can be used to reliably predict the identities of the track nodes and provide guidance to track nodes lacking the face modality. Nevertheless, this method involves several parameters, such as the trade-off parameter between

face and clothing modalities and the parameter to adjust the similarity in the non-local cost aggregation. These parameters may depend on the characteristics of the datasets. One future work could be the learning of the parameters so that the proposed method becomes adaptable to the characteristics of input data.

6.2 Ambiguously labeled learning

We introduced a matrix completion framework for resolving label ambiguities. The proposed method ensures all the instances and their associated ambiguous labels are utilized as a whole for resolving the ambiguity by discovering the underlying low-rank structure of subjects. Besides, we showed that the issue of labeling imbalance can be handled by performing column-wise weighting on the heterogeneous matrix. Moreover, the proposed iterative candidate elimination step can further improve the performance by iteratively removing the least confident candidate. Our method currently assumes instances of the same subject are jointly low-rank to resolve label ambiguity. This assumption can be violated if the feature vectors of a subject are selected from different domains or distributions. For instance, face images of different poses cannot be well approximated by a low-rank structure. One can generalize MCar to operate on instances of different distributions.

6.3 Target face association

We proposed a target face association method to retrieve a set of face images in the video using a single target annotation of the target face. These associated

face images provide a diverse representation and can be utilized to create a robust representation for video-based face identification. There are several directions in which this work can be extended. Currently, we use the average feature vectors of the associated face images to construct the robust representation. Nevertheless, the outliers in the associated face images can cause performance degradation. One interesting extension is to design a weighting scheme for constructing the robust representation by reducing the contribution from face images of lower association confidence. Besides, the performance of TFA can be affected by the characteristics of a face detector. False positive error of face detection can lead to erroneous association of non-face images, and missed detection of face images will prevent TFA from associating the useful exemplars for creating a robust template. Since the video quality and the distribution of face images are different from video to video, the performance of TFA can be potentially improved with a video-specific confidence threshold for face detection.

6.4 Face recognition using an outdoor camera network

We discussed some recent techniques for face recognition in outdoor camera networks and presented our implementation of the real-time video surveillance system at the University of Maryland campus. Several features can be integrated on top of the current master and slave camera framework to form a hybrid camera network. Mobile cameras, such as cell phone cameras, body cameras, and cameras mounted on the drones, can be integrated into the system to opportunistically retrieve the

face images in the monitored area. Moreover, face association across multiple views should be studied in the hybrid camera networks to fully utilize the information from multiple cameras.

Bibliography

[1] L. An, M. Kafai, and B. Bhanu. Dynamic bayesian network for unconstrained face recognition in surveillance camera networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2):155–164, Jun. 2013.

[2] M. Du, A. C. Sankaranarayanan, and R. Chellappa. Robust face recognition from multi-view videos. *IEEE Transactions on Image Processing*, 23(3):1105–1117, Mar. 2014.

[3] J. Sang and C. Xu. Character-based movie summarization. In *ACM Multimedia*, 2010.

[4] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Storygraphs: Visualizing character interactions as a timeline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] M. Everingham J. Sivic and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *Image and Video Retrieval*, 2005.

[6] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *ACM International Conference on Multimedia Retrieval*, 2014.

[7] Q. Yang. A non-local cost aggregation method for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[9] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic person identification in TV-series. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *European Conference on Computer Vision (ECCV)*, 2014.

[12] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[13] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[15] R. Ahuja, T. Magnanti, and J. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[16] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[17] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[18] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[19] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, Oct. 2007.

[20] M. Bäuml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy - Automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006.

[22] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Improved weak labels using contextual cues for person identification in videos. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[23] M. Du and R. Chellappa. Face association across unconstrained video frames using conditional random fields. In *European Conference on Computer Vision (ECCV)*, 2012.

[24] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture? In *Neural Information Processing Systems (NIPS)*, 2004.

[25] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[26] E. Hüllermeier and J. Beringer. Learning from ambiguously labeled examples. In *Intelligent Data Analysis*, 2006.

[27] L.-P. Liu and T. G. Dietterich. A conditional multinomial mixture model for superset label learning. In *Neural Information Processing Systems (NIPS)*, 2012.

[28] Y.-C. Chen, V. M. Patel, J. K. Pillai, R. Chellappa, and P. J. Phillips. Dictionary learning from ambiguously labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[29] L.-P. Liu and T. G. Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning (ICML)*, 2014.

[30] A. B. Goldberg, X. Zhu, B. Recht, J.-M. Xu, and R. D. Nowak. Transduction with matrix completion: Three birds with one stone. In *Neural Information Processing Systems (NIPS)*, 2010.

[31] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *Neural Information Processing Systems (NIPS)*, 2011.

[32] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, Sep. 2009.

[33] Z. Zeng, S. Xiao, K. Jia, T.-H. Chan, S. Gao, D. Xu, and Y. Ma. Learning by associating ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[34] Y.-C. Chen, V. M. Patel, R. Chellappa, and P. J. Phillips. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security*, 9(12):2076–2088, Dec. 2014.

[35] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, Jun. 2011.

[36] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[37] D. Huang, R. S. Cabral, and F. De la Torre. Robust regression. In *European Conference on Computer Vision (ECCV)*, 2012.

[38] J. Luo and F. Orabona. Learning from candidate labeling sets. In *Neural Information Processing Systems (NIPS)*, 2010.

[39] C. Ambroise, T. Denoeux, G. Govaert, and P. Smets. Learning from an imprecise teacher: Probabilistic and evidential approaches. *Applied Stochastic Models and Data Analysis*, 1:100–105, 2001.

[40] R. Jin and Z. Ghahramani. Learning with multiple labels. In *Neural Information Processing Systems (NIPS)*, 2002.

[41] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[42] T. Zhang. Statistical analysis of some multi-category large margin cassification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

[43] J. Cid-Sueiro. Proper losses for learning from partial labels. In *Neural Information Processing Systems (NIPS)*, 2012.

[44] M.-L. Zhang. Disambiguation-free partial label learning. In *SIAM International Conference on Data Mining*, 2014.

[45] M.-L. Zhang and F. Yu. Solving the partial label learning problem: An instance-based approach. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[46] A. Shrivastava, V. M. Patel, and R. Chellappa. Non-linear dictionary learning with partially labeled data. *Pattern Recognition*, 48(11):3283–3292, Nov. 2015.

[47] S. Xiao, D. Xu, and J. Wu. Automatic face naming by learning discriminative affinity matrices from weakly labeled images. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2440–2452, Oct. 2015.

[48] M. Sahare and H. Gupta. A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(3):160–164, 2012.

[49] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu. Towards class-imbalance aware multi-label learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[50] K. Chen, B.-L. Lu, and J. T. Kwok. Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In *International Joint Conference on Neural Networks (IJCNN)*, 2006.

[51] B.-L. Lu, K.-A. Wang, M. Utiyama, and H. Isahara. A part-versus-part method for massively parallel training of support vector machines. In *International Joint Conference on Neural Networks (IJCNN)*, 2004.

[52] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, Sep. 2015.

[53] B. Wu, S. Lyu, and B. Ghanem. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *AAAI Conference on Artificial Intelligence*, 2016.

[54] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):121–135, Jan. 2015.

[55] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, Jun. 1965.

[56] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In *International Joint Conference on AI*, 1999.

[57] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46(1).

[58] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215, November 2009*.

[59] E. J. Candès and B. Recht. Exact low-rank matrix completion via convex optimization. In *Allerton Conference on Communication, Control, and Computing*, 2008.

[60] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, Mar. 2004.

[61] S. Osher J.-F. Cai. Fast singular value thresholding without singular value decomposition. *UCLA CAM Report, May 2010.*

[62] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[63] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *European Conference on Computer Vision (ECCV)*, 2010.

[64] Z. Zhou and M. Zhang. Multi-instance multilabel learning with application to scene classification. In *Neural Information Processing Systems (NIPS)*, 2006.

[65] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the L1-ball for learning in high dimensions. In *International Conference on Machine Learning*, 2008.

[66] C.-H. Chen, V. M. Patel, and R. Chellappa. Matrix completion for resolving label ambiguity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[67] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *Proceedings of the 4th International Conference on Image and Video Retrieval*, 2005.

[68] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005.

[69] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-PEP for video face recognition. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[70] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[71] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[72] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[73] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016.

[74] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Y. Lee, V. E. Liong, J. W. Lu, M. de Assis Angeloni, T. de Freitas Pereira, H. X. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The IJCB 2014 PaSC video face and person recognition competition. *International Joint Conference on Biometrics (IJCB)*, 2014.

[75] L. Liu, L. Zhang, H. Liu, and S. Yan. Toward large-population face identification in unconstrained videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(11):1874–1884, Nov. 2014.

[76] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, Nov. 2004.

[77] M. Roth, M. Bauml, R. Nevatia, and R. Stiefelhagen. Robust multi-pose face tracking by multi-stage tracklet association. In *International Conference on Pattern Recognition (ICPR)*, 2012.

[78] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal. Online multi-face detection and tracking using detector confidence and structured SVMs. In *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2015.

[79] L. Ballan, M. Bertini, A. D. Bimbo, and W. Nunziati. Automatic detection and recognition of players in soccer videos. In *Proceedings of International Conference on Visual Information Systems*, 2007.

[80] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[81] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[82] M. Tapaswi, O. M. Parkhi, E. Rahtu, E. Sommerlade, R. Stiefelhagen, and A. Zisserman. Total cluster: A person agnostic clustering method for broadcast videos. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, 2014.

[83] S. Zhang, Y. Gong, J.-B. Huang, J. Lim, J. Wang, N. Ahuja, and M.-H. Yang. Tracking persons-of-interest via adaptive discriminative features. In *European Conference on Computer Vision (ECCV)*, 2016.

[84] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Joint face representation adaptation and clustering in videos. In *European Conference on Computer Vision (ECCV)*, 2016.

[85] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[86] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[87] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[88] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[89] B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo, T. D. Ngo, D.-D. Le, S. Satoh, and W. H. Hsu. Scalable face track retrieval in video archives using bag-of-faces sparse representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[90] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[91] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[92] J.-C. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, C.-H. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. *arXiv preprint arXiv:1605.02686*, 2016.

[93] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa. Triplet Probabilistic Embedding for Face Verification and Clustering. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2016.

[94] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, Dec. 2006.

[95] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

[96] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010.

[97] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, Mar. 2015.

[98] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, Sep. 2009.

[99] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2D still-image face recognition algorithms. *NIST Interagency Report 7709*, 2010.

[100] R. Ranjan, V. M. Patel, and R. Chellappa. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[101] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, Jun. 2008.

[102] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[103] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *European Conference on Computer Vision (ECCV)*, 2006.

[104] R. Olfati-Saber and N. F. Sandell. Distributed tracking in sensor networks with limited sensing range. In *American Control Conference*, 2008.

[105] C. Soto, B. Song, and A. K. Roy-Chowdhury. Distributed multi-target tracking in a self-configuring camera network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[106] F. Cupillard, F. Bremond, and M. Thonnat. Group behavior recognition with multiple cameras. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2002.

[107] M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[108] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[109] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer Vision (ECCV)*, 2014.

[110] V. Kulathumani, S. Parupati, A. Ross, and R. Jillela. Collaborative face recognition using a network of embedded cameras. *Distributed Video Sensor Networks, Springer*, pages 373–387, 2011.

[111] C. Ding, A. A. Morye, J. A. Farrell, and A. K. Roy-Chowdhury. Opportunistic sensing in a distributed PTZ camera network. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2012.

[112] C. Ding, B. Song, A. Morye, J. A. Farrell, and A. K. Roy-Chowdhury. Collaborative sensing in a distributed PTZ camera network. *IEEE Transactions on Image Processing*, 21(7):3282–3295, Jul. 2012.

[113] Y. Cai, G. Medioni, and T. B. Dinh. Towards a practical PTZ face detection and tracking system. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013.

[114] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. In *Proceedings of the IEEE*, 98(8):1485–1508, Aug. 2010.

[115] J. Zhao, S.-C. Cheung, and T. Nguyen. Optimal camera network configurations for visual tagging. *IEEE Journal of Selected Topics in Signal Processing*, 2(4):464–479, Aug. 2008.

[116] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, Jan. 1988.

[117] C. Piciarelli, L. Esterle, A. Khan, B. Rinner, and G. Foresti. Dynamic reconfiguration in camera networks: A short survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[118] M. Du and R. Chellappa. Face association for videos using conditional random fields and max-margin Markov networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1762–1773, Sep. 2016.

[119] X. Chen, J. Yang, and A. Waibel. Calibration of a hybrid camera network. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[120] G. Medioni, J. Choi, C.-H. Kuo, and D. Fidaleo. Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 39(1):12–24, Jan. 2009.

[121] V. M. Patel, J. Ni, and R. Chellappa. Remote identification of faces. In *Signal and Image Processing for Biometrics: State of the Art and Recent Advances, Springer*, 2014.

[122] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[123] X. Zou, B. Bhanu, B. Song, and A. K. Roy-Chowdhury. Determining topology in a distributed camera network. In *IEEE International Conference on Image Processing (ICIP)*, 2007.

[124] G. Medioni and Y. Cai. Persistent people tracking and face capture over a wide area. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[125] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[126] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[127] S. Duffner and J. Odobez. Track creation and deletion framework for long-term online multiface tracking. *IEEE Transactions on Image Processing*, 22(1):272–285, Jan. 2013.

[128] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[129] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *IEEE International Conference on Computer Vision workshop (ICCVW) on ChaLearn Looking at People (ChaLearn LaP)*, 2015.

[130] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994.

[131] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[132] D. Chen, X. D. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[133] D. Chen, X. D. Cao, L. W. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision (ECCV)*, 2012.

[134] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[135] H. Li and G. Hua. Hierarchical-PEP model for real-world face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[136] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, May 2013.

[137] S. Zhou, V. Krueger, and R. Chellappa. Face recognition from video: A CONDENSATION approach. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2002.

[138] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, Nov. 2004.

[139] Y.-C. Chen, V. M. Patel, P. J. Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In *European Conference on Computer Vision (ECCV)*, 2012.

[140] M. Marques and J. Costeira. 3D face recognition from multiple images: A shape-from-motion approach. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2008.

[141] L. Zhang, D. V. Kalashnikov, S. Mehrotra, and R. Vaisenberg. Context-based person identification framework for smart video surveillance. *Machine Vision and Applications*, 25(7):1711–1725, Aug. 2013.

[142] Z. Liu and S. Sarkar. Outdoor recognition at a distance by fusing gait and face. *Journal Image and Vision Computing*, 25(6):817–832, Jun. 2007.

[143] C. Pagano, E. Granger, R. Sabourin, G. L. Marcialis, and F. Roli. Adaptive ensembles for face recognition in changing video surveillance environments. *Information Sciences*, 286:75–101, Dec. 2014.

[144] F. Li and H. Wechsler. Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1686–1697, Nov. 2005.

[145] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, Jul 2013.

[146] T. B. Dinh, N. Vo, and G. Medioni. High resolution face sequences from a PTZ network camera. In *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, 2011.

[147] T. Dinh, Q. Yu, and G. Medioni. Real time tracking using an active pan-tilt-zoom network camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009.

[148] A. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *European Conference on Computer Vision (ECCV)*, 2000.

[149] Z. Wu and R. J. Radke. Keeping a Pan-Tilt-Zoom camera calibrated. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1994–2007, Aug. 2013.

[150] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *IEEE International Conference on Computer Vision (ICCV)*, 1999.

[151] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Second ed. Cambridge University Press, 2004.

[152] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: Exploring image collections in 3D. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2006)*, 2006.

[153] OpenCV. Open Source Computer Vision Library. http://opencv.org.

[154] Joo C. Neves, Juan C. Moreno, Silvio Barra, and Hugo Proen**c**a. Acquiring high-resolution face images in outdoor environments: A master-slave calibration algorithm. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015.

[155] A. A. Morye, C. Ding, A. K. Roy-Chowdhury, and J. A. Farrell. Distributed constrained optimization for bayesian opportunistic visual sensing. *IEEE Transactions on Control Systems Technology*, 22(6):2302–2318, Nov. 2014.

[156] Q. Lin, J. Yang, N. Ye, R. Wang, and B. Zhang. Face recognition in mobile wireless sensor networks. *International Journal of Distributed Sensor Networks*, 2013, 2013.

[157] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi. Privacy in mini-drone based video surveillance. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015.