# Discovering Person Identity via Large-Scale Observations

Yongkang Wong[1], Lekha Chaisorn[1], Mohan S. Kankanhalli[1,2]

[1]Interactive & Digital Media Institute, National University of Singapore, SG
[2]School of Computing, National University of Singapore, SG

**Abstract.** Person identification is a well studied problem in the last two decades. In a typical automated person identification scenario, the system always contains the prior knowledge, either person-based model or reference mugshot, of the person-of-interest. However, the challenge of automated person identification would increase by multiple folds if the prior information is not available. In today's world, rich and large quantity of information are easily attainable through the Internet or closed-loop surveillance network. This provides us an opportunity to employ an automated approach to perform person identification with minimum prior knowledge, presume that there are sufficient amount of observations. In this paper, we propose a dominant set based person identification framework to learn the identity of a person through large-scale observations, where each observation contains instances from various modality. Through experiments on two challenging face datasets we show the potential of the proposed approach. We also explore the conditions required to obtain satisfy performance and discuss the potential future research directions.

## 1 Introduction

Today we are living in a world of big data. With the recent advance in hardware technology, telecommunication protocol, and the growing popularity of social media, we are blessed with the rich and large quantity of data that are easily attainable through the Internet or closed-loop surveillance network. However, while it is relatively easy, albeit expensive, to install servers to handle the increasing demand on storage and computational performance, it is quite another issue to adequately monitor and analyze the data, big data. On the other hand, big data has provided the research community new research opportunities and directions. Therefore, the question is how can we utilize the big data for better innovations.

In the last two decades, person identification has received a lot of attention from the computer vision and machine learning community [1–6]. For example, iris recognition [1], fingerprint recognition [2], face recognition [3–5], and gait recognition [6]. In the literature, any of the person identification problem can be generalized into three distinct configurations: closed-set identification, open-set identification, and verification [7]. The task of closed-set identification is to

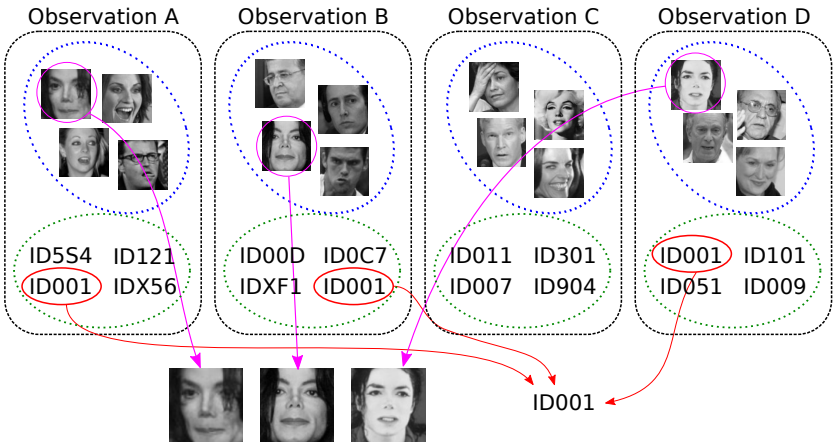


Fig. 1: Conceptual example of instance association based person identification with multiple observations. Given four unique observations, where each observation contains a number of instances of the "personal ID" and "facial image". We can analyze the IDs to select a subset of salient observations (i.e., A, B and D) for ID0001, followed by identify the candidate images with high internal coherency. Note that the relationship between the identity labels and images are unknown.

classify a given face as belonging to one of the $K$ previously seen persons in a gallery, where as the open-set identification take into account the possibilities of impostor attack with an additional "unknown person" class. The task of verification is to determine if two given samples belong to the same individual, where one or both identities may not have been observed beforehand [3].

In this work, we aim to address a different category of identification problem, namely *instance association* based person identification, where the identity of each previously unseen individual is learned through single observation [8, 9] or a large number of observations. The identification problem on hand is a weakly label learning problem, where each observation contains multiple instances and labels without given relationships. For example, assume that we can access to a large number of articles (*observations*) and each article contains multiple facial images and name-entities (*instances*). If there exist a genuine subset of salient observations, such that these observations all contain name-entities from a unique individual (denoted as salient ID) and no other name name-entities co-occur in the exact set of observations, it is intuitive to assume that the same facial image that co-occurs in these observations represent the salient ID. Under this scenario, the identification task can only be performed if enough samples are observed at several observations. Note that if the association between a name-entity and the correspond facial image is provided as prior knowledge, i.e., a facial image and its genuine identity are given, it will be classified as open/closed-set identification problem. A conceptual example of instance association based identification problem with multiple observations is shown in Figure 1.

Here, we illustrate two real-world surveillance applications. Assume that we are provided with video feeds from Closed-Circuit Television (CCTV), as well as the datalog from the wireless access points and access control system, we can form a large collection of spatial-temporal observations over time, where each observation contains video footage and electronic signals appear at a specific location and duration. In an event of criminal act, a security officer can shortlist a number of candidates (e.g., electronic signal detected at point A), then the system can analyze the achieved observations to short-list the visual images of these candidates. Another practical example is to associate the visual image of the holder of a stolen access card. It is important to note that CCTV generally does not cover all the area. Hence, the identification on hand would need to consider the problem of incomplete observations from visual data. We will address this realistic scenario in the future work.

In this work, we proposed a dominant set based person identification framework to learn the identity of a person through multiple observations, where each observation contains instances from two modalities (i.e., name-entity and facial image). We consider the learning problem as a graph labeling problem, where the instances in each observation are considered as a vertex set of graph. By using an intuitive salient observations detection stage, the problem of graph labeling can be cast as a dominant set clustering problem. Given the dominant set clustering output and consider the structure of observations we proposed three instance selection approaches to perform person identification. The proposed framework is quantitatively and qualitatively evaluated on two challenging face datasets, where one of the dataset contains large number of face images obtained from the Internet. To understand the limitation of this problem, we simulated a number of variation of co-occurrence rate in the salient observations.

**Contributions.** In this work, we describe a person identification problem, namely *instance association based person identification*, which assume the identity of a person is not directly given but available in a weakly label manners. The described problem is realistic and applicable to several surveillance applications. We propose to use dominant set based approach to addressed this identification task and mathematically outline the problem as a graph labeling problem. Experimental results on two challenging face datasets under various constraints shows the potential of the proposed approach.

We continue the paper as follows. Section 2 describes related work. We formulate the problem of instance association based person identification in Section 3, where the proposed framework is presented and discussed in Section 4. Section 5 is devoted to experiments and Section 6 provides the main findings and future directions.

## 2 Related Work

Person identification, or identity inference, has received significant attentions over the last two decades. Generally, the person identification task always assumes the mugshots of person-of-interest are always known, where the task of

identification is to either classify a given face as belonging to one of the previously seen persons in a gallery (i.e., closed-set identification), or belonging to an unknown individual (i.e., open-set identification) [7]. In the literature, there exist very little work to perform person identification without the gallery of persons. One related work is to associate people appearing in news video with their names [8–10]. The task of this work is to find the video segments where a person appears and associate the person with a name. Satoh *et al.* [8] explored the co-occurrence of facial images, video caption, and transcript in a news video. The underlying idea is that the (similar) faces that frequently co-occur with a certain name are likely to match the name and vice versa. Houghton [10] designed an automated system to create a named faces database, which utilize a similar approach as [8] to analyze the content on websites and news videos. Yang and Hauptmann explored the *features* and *constraints*, which reveal the relationships among the names of different people, to perform name association in news video [9]. We note that the aforementioned problem is focus on a single observation scenario. In addition, the validity of co-occurrence might not whole for all observation. For example, a news video can contains many faces without mentioned all the names, or without any facial images of the identity of interest. In contrast, the assumption of co-occurrence is more likely to hold with large number of observations, ideally on a scale of big data. However, the difficulty of the learning problem is also increased dramatically.

Different from the names and face association application, Cho *et al.* [11, 12] proposed a system to dynamically associate the personal identification, obtained via RFID system, fingerprint, iris recognition, etc., to the persons observed by the visual sensors. The system assumed there exists a gate region to collect the personal identity and each individual is continuously detected and tracked in the premises. In this system, the association is achieved with heterogeneous sensors of visual sensors and identification sensors, which required carefully calibrated information to associate the identity to the individual in a particular location. Our proposed problem does not have this constraint.

The instance association based person identification problem is related to Multi-Instance Multi-Label (MIML) learning problem [13, 14]. Under the formulation of MIML [13], each object is described by multiple instances and associated with multiple class labels. The goal is to learn a classifier to classify the genuine instance for each label. This formulation is very useful for complex object with multiple semantic meaning, such as scene recognition, text categories classification, etc. In the domain of this person identification problem, the instances (a.k.a. personal attribute) can be extracted from various modality. Specifically, the personal attributes can be categorized into biometrics attributes or non-biometrics attribute. The biometric attribute include gait, iris, facial image, fingerprint, hand-geometry, etc., where the non-biometric attribute are name, age, contact information (such as email, phone number, etc.), and personal affiliation. For future research direction, this person identification learning problem can be modeled as a multi-modality instances learning problem. We will address this in future work.

# 3 Problem Formulation

## 3.1 Instance Association based Person Identification via Graph Labeling

In the context of big data, we assume that each individual can be observed with a variety of sensors. Through the observations from these sensors, the data can be modeled as a super-set of collection $\mathbb{O}$, which comprises of a finite number of local observations $\boldsymbol{O}^m$ for $m = 1, 2, \ldots, M$. Each of the $m$-th observation consist of a collection of instances from a specific spatial-temporal subspace. In addition, the instances collected in each observation belong to a dedicated modality and the relationship between the instances are unknown. For example, $\boldsymbol{O}^m$ can consists of all the faces and audio recording segments collected over a short period of time at hotel lobby, or the faces and names from a newspaper.

We first consider the graph labeling problem using instances from a single modality. For example, gait, iris, facial image, etc. Hence, each observation $\boldsymbol{O}^m$ consists of one bag of instance $\mathbb{X}^{m,s} = \{\boldsymbol{x}_1^{m,s}, \boldsymbol{x}_2^{m,s}, \ldots, \boldsymbol{x}_N^{m,s}\}$, where $\boldsymbol{x} \in \mathbb{R}^d$ is a $d$-dimensional vector from modality $s$. Each instance is considered as a vertex and connected to its neighbors through undirected edges that having positive weights. Given a local observation set $\mathbb{O} = \{\boldsymbol{O}^1, \boldsymbol{O}^2, \ldots, \boldsymbol{O}^M\}$, the task of a graph label $\varphi : \mathbb{O} \mapsto \widehat{\mathcal{X}}$ is to produce $\widehat{\mathcal{X}} = \{\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2, \ldots, \widehat{\boldsymbol{x}}_c\}$, where ideally instances in $\widehat{\mathcal{X}}$ belong to the same target $\mathcal{T}$.

Now, lets consider the scenario of multi-modality graph labeling problem. Here, the $m$-th observation is re-written as $\boldsymbol{O}^m = \{\mathbb{X}^{m,1}, \mathbb{X}^{m,2}, \ldots, \mathbb{X}^{m,S}\}$ where $\mathbb{X}^{m,s}$ is a bag of instances from modality $s$. Follow the scenario of the single modality approach, the task is to employ a graph label function $\varphi : \mathbb{O} \mapsto \widehat{\mathcal{X}}$ to produce $\widehat{\mathcal{X}}$. Differing from the single modality scenario, we are now facing the problem of matching instances from different modality. Note that the latent relationship of instances from different modality may be unknown in our problem, e.g., a voice pattern may not be associated to any face images in an observation. In this work, we formulate the multi-modality person identification problem as iterative multi-stage clustering problem, where each iteration is dedicated to identify a randomly selected individual. Specifically, we perform the following task in each iteration. Given the mega set of observations $\mathbb{O}$, we employ a salient observation classifier $\mathcal{F} : \mathbb{O} \mapsto \mathbb{O}^l$ to produce $\mathbb{O}^l$, where each observation in $\mathbb{O}^l$ contains at least one instance that belong to target $\mathcal{T}^l$. Note that $\mathbb{O}^l$ can be extracted with non-biometrics instances, such as electronics signal or name-entity, which give high confidence decision when compared to biometrics instances (e.g., facial images or gait). Given the newly extracted $\mathbb{O}^l$, the graph labeling problem is now become a dominant-set detection problem, where the dominant set can be extract with $\varphi_{\text{DS}} : \mathbb{O}^l \mapsto \widehat{\mathcal{X}}$.

# 4 Proposed Method

In this section, we elaborate the proposed framework for the instance association based person identification problem. In this work, we evaluate the propose

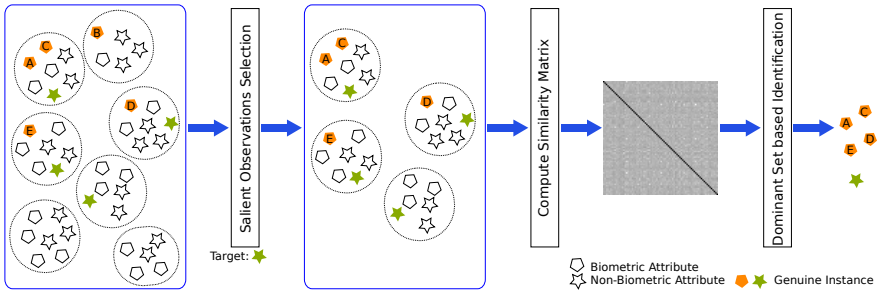


Fig. 2: Conceptual diagram of the proposed instance association based person identification framework. In this example, the system is assigned to perform person identification for a particular target.

framework with two types of personal attribute, i.e., facial image (biometric attribute) and person ID (non-biometric attribute), where the generalization to other types of modality is discussed in Section 6. To reduce the complexity of this work, we assume that the face image and ID of the same individual will always co-occur in the same observation, where the scenario of missing data will be addressed in future work. The proposed framework is an iterative person identification process and each iteration comprised of three components: **(1)** salient observations detection, **(2)** compute similarity matrix with biometrics instances, and **(3)** person identification with dominant set analysis. Given a finite set of observations, $\mathbb{O}$, the task of the proposed framework is to iteratively analyze each of the available person IDs, and perform person identification task if the detection criteria (see Section 4.1) is satisfied. A conceptual overview of the proposed method is shown in Figure 2.

### 4.1 Salient Observations Detection

Given a finite set of observations, $\mathbb{O}$, and an ID-of-interest (IDoI), $l$, the task is to find a subset of salient observations, $\mathbb{O}^l$, which satisfy two criteria: **(1)** all observations should contain at least one instance of IDoI, and **(2)** the IDs that not belong to IDoI should not be observed in the same set of observations. This is to ensure that each salient observations has at least one genuine face image from each of the salient observations. Given the detected $\mathbb{O}^l$, we refer to the face images that belong to IDoI as genuine and the remaining images as imposter.

To measure the quality of the detected $\mathbb{O}^l$ with respect to the person identification problem, we define two properties, namely *observability* and *observation noise*. The *observability* is a binary attribute which combines the two aforementioned criteria. If the observability of a $l$ is false, this indicate that $\mathbb{O}^l$ can not be used to find the associated genuine image of $l$. This generally happen if IDoI only appear in one observation or there exist other ID(s) that co-occurs in the same set of observations. The *observation noise*, denoted as `ON`, is a scalar attribute that correlate the rate of an imposter ID that co-occur in $\mathbb{O}^l$. $\texttt{ON} = 1.0$ indicates there exist an imposter that co-occur in all observations in $\mathbb{O}^l$. Lower value of ON lead to better chance of successful person identification.

### 4.2 Similarity Matrix with Face Verification

Given the IDoI $l$ and the extracted salient observations $\mathbb{O}^l = \{\boldsymbol{O}_1^l, \boldsymbol{O}_2^l, \ldots, \boldsymbol{O}_M^l\}$ where the $m$-th observation has $|\boldsymbol{O}_m^l|$ face images, we can represent the face images as an undirected edge-weighted (similarity) graph with no self-loops $G = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{w})$. Here, $\boldsymbol{V} = \{1, \ldots, n\}$ is the vertex set, $\boldsymbol{E} \subseteq \boldsymbol{V} \times \boldsymbol{V}$ is the edge set, and $\boldsymbol{w} : E \to \mathbb{R}_+^*$ is the (positive) weight function. We define a $N_{\mathbb{O}} \times N_{\mathbb{O}}$ symmetric matrix $\boldsymbol{W}$ of pairwise similarities between candidate face images, where $N_{\mathbb{O}} = \sum_{m=1}^{M} |\boldsymbol{O}_m^l|$ is the total number of face images in the salient observations. Using the symmetric similarity matrix, the goal is to perform graph-based clustering method (see Section 4.3) to retrieve a dominant set of instances, where the dominant instances form the most coherent subset.

**Image Representation:** In this work, we employ a local feature-based face representation, namely Locally Sparse Encoded Descriptor (LSED), which has shown good robustness against various alignment errors and pose mismatches with various face dataset, as well as its simplicity in implementation [3]. Briefly, a given face image is first split into $R$ fixed size regions, followed by a secondary split into small overlapping blocks with a size of $8 \times 8$ pixels. Each block is represented by a low-dimensional texture descriptor, $\boldsymbol{y}$, followed by a sparse coding based encoding method to obtain a block level sparse descriptor. The $r$-th region descriptor, $\boldsymbol{h}_r$, is computed by pooling the block level sparse descriptor from region $r$ with average pooling operation. Due to the relaxed spatial constraints within each region, it allows some movement and/or deformations of the face components and leads to a degree of inherent robustness to expression and pose changes [15, 3]. In this work, we select the implicit sparse encoding via probabilistic approach in [3] due to its robust performance under various image conditions and light weight computational cost. The $i$-th block level sparse descriptor in region $r$ can now be computed via:

$$\boldsymbol{h}_{r,i} = \left[ \frac{w_1 p_1\left(\boldsymbol{y}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{y}_{r,i}\right)}, \cdots, \frac{w_1 p_G\left(\boldsymbol{y}_{r,i}\right)}{\sum_{g=1}^{G} w_g p_g\left(\boldsymbol{y}_{r,i}\right)} \right]^T \tag{1}$$

where the $g$-th element in $\boldsymbol{h}_{r,i}$ is the posterior probability of $\boldsymbol{y}_{r,i}$ according to the $g$-th components of a Gaussian Mixture Models, and $w$ is the associated weight. We direct the user to [3] for details descriptions. A conceptual example of LSED face descriptor can be found on Figure 3.

**Similarity Matrix:** Considering a pair of face images $A$ and $B$, the similarity is then defined as the inverse of a cohort normalization [16] based distance, written as

$$w_{A,B} = \frac{1}{d_{\texttt{norm}}(A, B)} \tag{2}$$

where

$$d_{\texttt{norm}}(A, B) = \frac{d_{\texttt{raw}}(A, B)}{\sum_{i=1}^{N_C} s_{\texttt{raw}}(A, C_i) + \sum_{i=1}^{N_C} s_{\texttt{raw}}(B, C_i)} \tag{3}$$

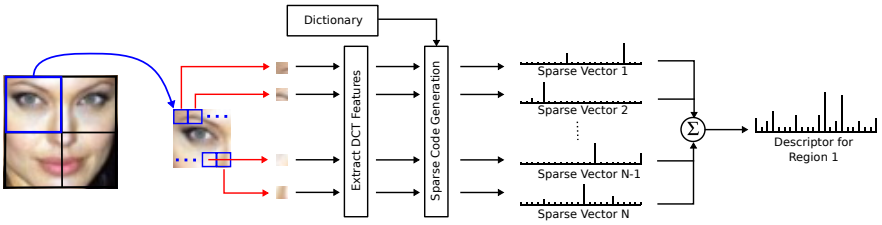


Fig. 3: Conceptual example of Locally Sparse Encoded Descriptor for face images. The local image patches are encoded locally via sparse representation based encoding, where the region level pooling enforce the regional structure information of face images.

The cohort faces $C_i$ are assumed to be reference faces that are different form the images of persons $A$ or $B$. The raw distance, $d_{\texttt{raw}}(A, B)$, is obtained by compare the corresponding face regions of the two images using Euclidean distance.

### 4.3 Dominant Set based Person Identification

Based on the discussion from Section 4.1, it is intuitive that each observation in $\mathbb{O}^l$ contains at least one instance of face image that belong to the IDoI $l$. Therefore, the person identification problem can be re-casted as a dominant set clustering problem. Following [17], the cluster of vertices can be associated to a $N_{\mathbb{O}}$-dimensional vector, $\boldsymbol{x}_{\texttt{DS}}$, where its components express the *participation* of nodes in the cluster. Intuitively, the dominant components in $\mathbb{O}^l$ should be belong to the IDoI $l$, where the associated components will have large value in $\boldsymbol{x}_{\texttt{DS}}$. One way to define the *cohesiveness* of a cluster is given by the following quadratic form [17]:

$$f(\boldsymbol{x}_{\texttt{DS}}) = \boldsymbol{x}_{\texttt{DS}}^T \boldsymbol{A} \boldsymbol{x}_{\texttt{DS}} \tag{4}$$

where the element $i, j$ of $\boldsymbol{A}$, $a_{i,j}$, is equal to $w(i, j)$ if $(i, j) \in \boldsymbol{E}$. Note that $\boldsymbol{A}$ is an undirected graph with no self-loops, therefore all the element in the main diagonal of $\boldsymbol{A}$ is zero. Now, the dominant set clustering problem can be solve by finding an optimum vector $\boldsymbol{x}_{\texttt{DS}}$ to maximize $f$.

One way to find the optimal solution is with the *replicator equation* [18]. First, each element of $\boldsymbol{x}_{\texttt{DS}}$ is initialized to $1/|N_{\mathbb{O}}|$, followed by iteratively solve the following model:

$$\boldsymbol{x}_{\texttt{DS},i}(t+1) = \boldsymbol{x}_{\texttt{DS},i}(t) \frac{(\boldsymbol{A}\boldsymbol{x}_{\texttt{DS}})_i}{\boldsymbol{x}_{\texttt{DS}}(t)^T \boldsymbol{A} \boldsymbol{x}_{\texttt{DS}}(t)} \tag{5}$$

The algorithm terminates when $f\left(\boldsymbol{x}_{\texttt{DS}}(t+1)\right) - f\left(\boldsymbol{x}_{\texttt{DS}}(t)\right) < \epsilon$, where the parameter $\epsilon$ is the stopping criterion. In addition to the aforementioned termination criteria, the algorithm will be terminated after $T$ iterations.

Given the optimal $\boldsymbol{x}_{\texttt{DS}}$ that maximize Equation 4, we are able to perform human identification with three candidate selection approaches.

- *local observation analysis:* From Section 4.1, we know that each observation contains at least one genuine candidate. Therefore, the baseline approach is

to select the instances with the highest participation in their corresponding observation. The dilemma of this approach is that if the observations contain more than one genuine instance the recall will reduce. We term this approach *top selection approach.*

- *maximize internal coherence:* The selection of dominant candidates is conducted by selecting the instances with high participation to the dominant set cluster, which can be obtained if $x_{\mathtt{DS},i} > \tau$ and the threshold $\tau$ determine the strength of the participation. However, the the genuine instances will be ignored if the corresponding value in $\boldsymbol{x}_{\mathtt{DS}}$ is lower than $\tau$. This is particular obvious when the face image is exposed to multiple types of environment variations and pose changes. We note this could be a potential problem for this approach but can be solved with the improvement in face matching algorithm. We term this approach *threshold-based approach.*
- *fusion approach:* This approach combine the above selection approaches. Given the optimum $\boldsymbol{x}_{\mathtt{DS}}$, we first extract all candidates that satisfy $x_{\mathtt{DS},i} > \tau$, followed by select the instance with the highest participation in their corresponding observation if no instance is classified as genuine in threshold stage. We term this approach *fusion-based approach.*

## 5 Experiments

In this section, we examine the performance of the proposed framework for instance association based human identification problem. We first provide an overview of the image datasets and protocol used in the experiments, followed by the evaluation metrics. Then, we quantitatively analyze the performance under various configurations and provide qualitative comparisons.

### 5.1 Image Dataset and Protocol

Experiments were conducted on three datasets: Yale Face Dataset B [19], extended Yale Face Dataset B [20], and the Labeled Faces in the Wild (LFW) dataset [21]. The first two datasets are comprised of facial images captured in a laboratory configuration with various illumination variations, where the LFW dataset contains real-world facial images obtained by automatic crawling the Internet.

The Yale Face dataset B was explicitly created to study the face recognition performance under the influence of illumination variations. We combined the frontal view images of both Yale face dataset B, denoted as YaleB, and produce a total of 2,455 images (with 64 illuminations conditions) of 38 individuals. The cropped grayscale facial images were extracted with the manual labeled eye coordinates, and have the size of $64 \times 64$ pixels. Each of the images has zero degree in-plane rotation and the inter-ocular distance was 32 pixels (located at $(15, 19)$ and $(47, 19)$). This dataset was used to evaluate the proposed framework under small variation of illumination conditions and various level of $\mathtt{ON}$. The face images were divided into two sets: training set and evaluation set. We randomly

selected 8 images from each individual to form the evaluation set, where the remaining images were assigned to the training set. The training set is dedicated for LSED's dictionary training and cohort selection. For the evaluation set, we randomly created 10 set of observations for each individual with each observation contains 8 face images. In total, we generated 4 evaluation sets with different level of ON (i.e., 0.2, 0.4, 0.6, and 0.8)[1].

The LFW dataset was designed to provide a platform to study face recognition performance under uncontrolled environments. It contains 13,233 face images of 5,749 individuals. Among all the subjects, we selected 158 individual with 10 facial images of more as our genuine set and the remaining individual are used as impostors. In this work, we cropped the originally detected face images (i.e., without using additional algorithm to correct the alignment errors) by a fixed bounding box with coordinates $(62, 71)$ to $(187, 196)$ and rescale to $96 \times 96$ pixels. We used the training set from LFW view 1 for dictionary training and cohort selection. We created an evaluation set of 1,200 observations where each observation contains 8 instances of facial images and names. The evaluation set contains 158 subset of salient observations that fulfill the observability criterion and ON was limited to 0.2.

We qualitatively report the performance of the proposed algorithm with F-measure metric, which is

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

where the precision and recall are defined as $TP/(TP+FP)$ and $TP/(TP+FN)$, respectively. The notation $TP$, $FP$, and $FN$ are the total number of true positive, false positive, and false negative identification (computed with each set of salient observations), respectively. In addition, we also report the Receiver Operating Characteristic (ROC) curve, where the True Positive Rate ($TPR$), False Positive Rate ($FPR$) are defined as $TP/(TP+FN)$ and $FP/(FP+TN)$, respectively. The notation $FN$ is the total number of false negative identification.

### 5.2 Performance Evaluation

We evaluate the proposed method with three candidate selection approaches (see Section 4.3 for details), namely top selection approach, threshold-based approach, and fusion-based approach. The effect of selection threshold $\tau$ is discussed in this section. Based on preliminary experiments, the proposed approach used the following parameters: the face images are divided into $5 \times 5$ regions for YaleB dataset and $3 \times 3$ regions for LFW dataset. Dimension of each DCT-based texture descriptor is 15, and the number of visual words in the dictionary is 1024. The stopping criterion $\epsilon$ is set to 0.001.

In experiment 1, we first quantitatively evaluate the performance with F-measure over four configuration of ON. The evaluation is conducted on YaleB dataset as the facial images are well aligned and captured under strict controlled.

[1] The protocol will be publicly available

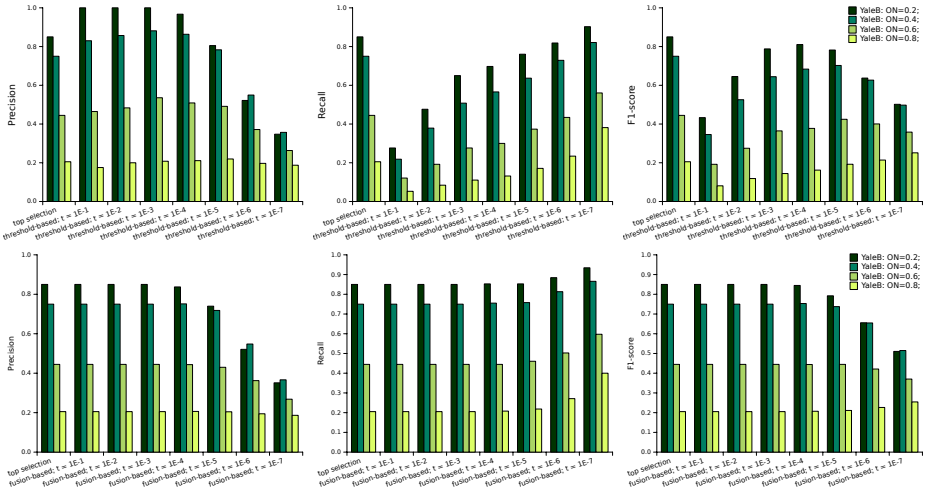


Fig. 4: Performance comparison of various threshold $\tau$ on the YaleB dataset. Top row is the performance of the threshold-based approach whereas the bottom is with the fusion-based approach. The first column in each plot are the performance with top selection approach.

The average precision, recall, and $F_1$-score are shown in Figure 4. As shown in the figure, the threshold-based approach achieved precision of 1.0 (`ON` = 0.2) but lower value in recall when $\tau$ is larger than 0.001. The best $F_1$-score is obtained when $\tau = 0.0001$. The similar performance pattern is observed across all variance of `ON`, where the performance with `ON` = 0.8 is considered as noise. We visually evaluate the output of the identification and found that the classified images are generally belong to the same impostor. This indicates that the influence of `ON` is signification and is the core challenge in our problem. For potential application in real-world deployment of such system, the automated system should avoid identification if the noise level is too high. For the fusion-based approach, the performance when $\tau$ is higher than 0.0001 is identical to the top selection approach. This is expected as the top participant is each observation is selected when no candidates satisfy the selection threshold. The inclusion of this clearly improve the performance. The performance is consistently better than the threshold-based approach in recall and $F_1$-score.

In addition to the F-measure metric, we compare the performance of threshold-based approach and fusion-based approach with ROC curve on both datasets. As shown in Figure 5, the performance reduced when we increased the observation noise `ON`. As expected, the performance with LFW dataset (`ON` was limited to 0.2) is lower than YaleB dataset. This is acceptable due to the variations in image quality and capture conditions. The qualitative comparison will be discussed in next section. The analysis also shows that the area under the ROC curve with fusion-based approach is generally higher than threshold-based approach. The only exception is when `ON` is equal to 0.8.

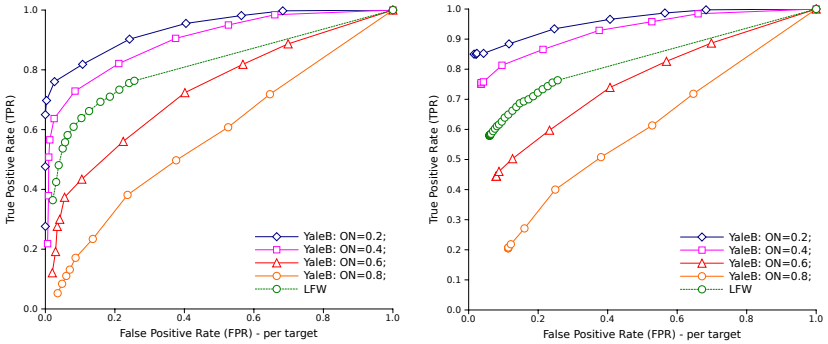


Fig. 5: ROC curve of person identification with YaleB and LFW dataset using: (*left*) threshold-based approach and (*right*) fusion-based approach.

In experiment 2, we qualitatively compare the performance of the proposed method on YaleB and LFW dataset, where the observation noise `ON` was limited to 0.2. We compare the identification output with the top selection approach and threshold-based approach (with a selected number of $\tau$). Results are shown in Figure 6 and Figure 7. Through the analysis on the YaleB dataset, we found that the optimum value of $\tau$ vary across different individual. Unlike the closed/open-set identification problem, it is impractical for us to tune this parameter as the image conditions and facial expression vary across different real-world application. We note that the best approach is to use a better face descriptor (or face matching algorithm) to stabilize the impact from these factors. A good similarity score normalization algorithm can be considered. Another observation we made is that the eye region of the individual change when the illumination conditions is different. For example, participant might close his eye when the flash level is high (shown in the last row of Figure 6). For the LFW dataset, we learned that the head gear (e.g., spectacle, hat, etc.) plays an important role in our evaluation (see the first results on Figure 7). We also select an example where all the predictions are wrong (see the bottom row in Figure 7). In this particular example, the prediction is heavily affected by the facial expression, which also affect to the most confidence selection scenario for the threshold-based approach (i.e., $\tau = 0.1$). Another possible future work is to employ fusion algorithm to utilize the strength of multiple algorithms. A detail study will be shown in future publications.

## 6 Main Findings and Future Directions

In this paper, we propose a novel problem for real-world person identification application, namely *instance association based person identification*, which is motivated with the increasing number of real-world data from the cyber-physical space and the improve accessibility to these data. Despite the large number of literature in person identification problem, most of them assume the mugshots and the identity associated information is given, where the goal is to identify if a

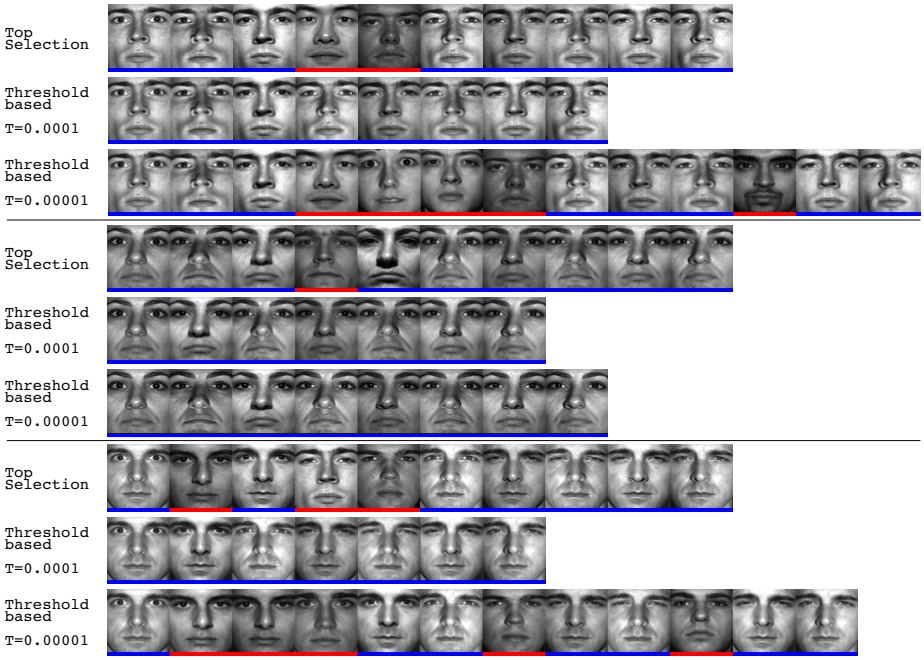


Fig. 6: Qualitative comparison of the proposed method with YaleB dataset. Blue and red indicate correct and incorrect identification, respectively.

current probe image belong to the known person or not. In this paper, we assume that this information is unknown and the goal is to learn the real identity of an image via large-number of observations. Formally, we divide the data into multiple spatial-temporal constrained observations, where each observation contains a finite number of instances from different modalities. We formulate the necessary components for the instance association based person identification problem. Through the observation of non-biometric attribute in the observations, we extracted the salient observations which satisfy the conditions to learn the genuine identity of an attribute-of-interest. We shown that the problem can be formulated as a dominant set clustering problem. Performance on two challenging face datasets, i.e., Yale dataset B and the Label Faces in the Wild (LFW) dataset, shows promising performance for the person identification problem on hand.

For future research directions, we would like to cast the person identification problem on hand as the multi-instance multi-label learning problem [13, 14]. In particularly, we would like to extend the existing work to simultaneously associate instances from various biometrics (i.e., multi-modality). Another research direction is to address the missing data problem (i.e., without the co-occurrence assumption) in the observations, this problem is deliberately ignored in this paper. Last but not least, we would like to emphasize that the proposed problem is practical and envisage the potential to apply this problem to the other real-world applications.

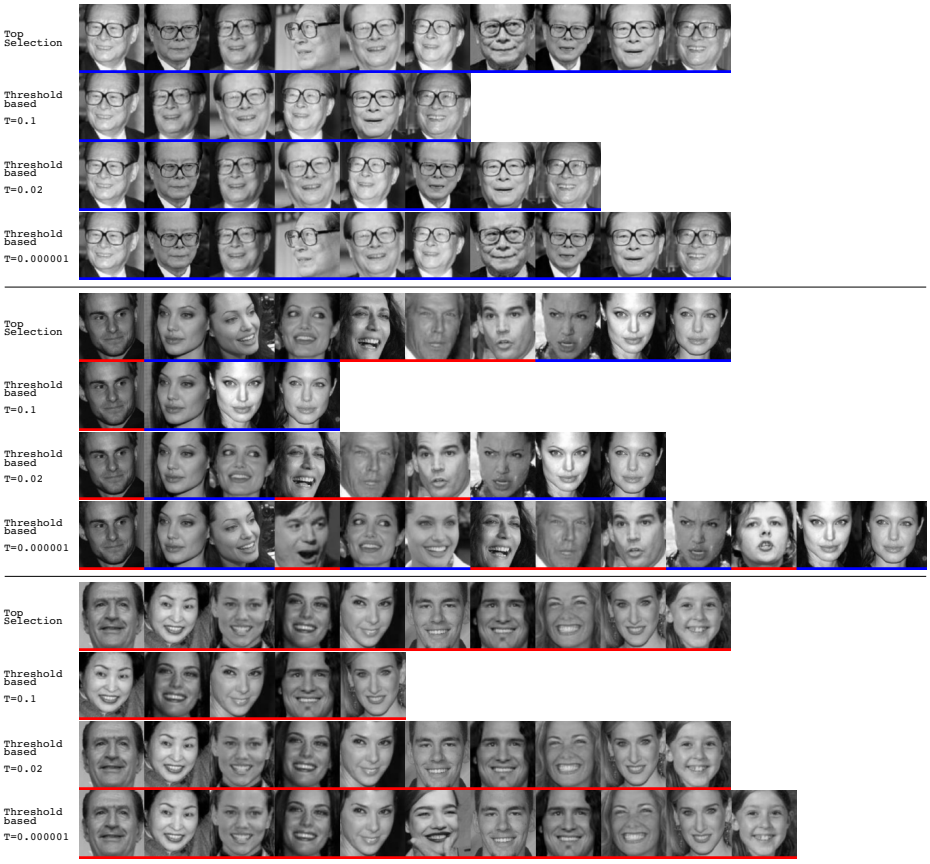


Fig. 7: Qualitative comparison of the proposed method with LFW dataset. Blue and red indicate correct and incorrect identification, respectively.

## 7 Acknowledgment

This research was carried out at the NUS-ZJU Sensor-Enhanced Social Media (SeSaMe) Centre. It is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office.

## References

1. Bowyer, K.W., Hollingsworth, K., Flynn, P.J.: A survey of iris biometrics research: 2008-2010. In: Handbook of Iris Recognition. Springer (2013) 15–54
2. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. 2 edn. Springer (2009)
3. Wong, Y., Harandi, M.T., Sanderson, C.: On robust face recognition via sparse coding: the good, the bad and the ugly. IET Biometrics ((in press))

4. Zhang, X., Gao, Y.: Face recognition across pose: A review. Pattern Recognition **42** (2009) 2876–2896
5. Zhao, W.Y., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys **35** (2003) 399–458
6. Liu, L.F., Jia, W., Zhu, Y.H.: Survey of gait recognition. In: Lecture Notes in Computer Science. Volume 5755. (2009) 652–659
7. Cardinaux, F., Sanderson, C., Bengio, S.: User authentication via adapted statistical models of face images. IEEE Transactions of Signal Processing **54** (2006) 361–373
8. Satoh, S., Nakamura, Y., Kanade, T.: Name-it: Naming and detecting faces in news videos. IEEE MultiMedia **6** (1999) 22–35
9. Yang, J., Hauptmann, A.G.: Naming every individual in news video monologues. In: Proceedings of ACM International Conference on Multimedia. (2004) 580–587
10. Houghton, R.: Named faces: Putting names to faces. IEEE Intelligent Systems **14** (1999) 45–50
11. Cho, S.H., Hong, S., Nam, Y.: Association and identification in heterogeneous sensors environment with coverage uncertainty. In: IEEE International Conference on Advanced Video and Signal Based Surveillance. (2009) 553–558
12. Cho, S.H., Hong, S., Moon, N., Park, P., Oh, S.J.: Object association and identification in heterogeneous sensors environment. EURASIP Journal on Advances in Singal Processing (2010)
13. Zhou, Z., Zhang, M., Huang, S., Li, Y.: Multi-instance multi-label learning. Artificial Intelligence **176** (2012) 2291–2320
14. Yang, S., Jiang, Y., Zhou, Z.: Multi-instance multi-label learning with weak label. In: International Joint Conference on Artificial Intelligence. (2013)
15. Heisele, B., Ho, P., Wu, J., Poggio, T.: Face recognition: component-based versus global approaches. Computer Vision and Image Understanding **91** (2003) 6–21
16. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. Speech Communication **31** (2000) 225–254
17. Pavan, M., Pelillo, M.: Dominant sets and pairwise clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 167–172
18. Weibull, J.W.: Evolutionary Game Theory. 1 edn. The MIT Press (1997)
19. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 643–660
20. Lee, K.C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 684–698
21. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)