Adaptive Learning for Celebrity Identification With Video Context

Chao Xiong, Guangyu Gao, Zhengjun Zha, *Member, IEEE*, Shuicheng Yan, *Senior Member, IEEE*, Huadong Ma, *Member, IEEE*, and Tae-Kyun Kim, *Member, IEEE*

***Abstract*—In this paper, we propose a novel semi-supervised learning strategy to address the problem of celebrity identification. The video context information is explored to facilitate the learning process based on the assumption that faces in the same video track share the same identity. Once a frame within a track is recognized confidently, the label can be propagated through the whole track, referred to as the *confident track*. More specifically, given a few static images and vast face videos, an initial weak classifier is trained and gradually evolves by iteratively promoting the confident tracks into the "labeled" set. The iterative selection process enriches the diversity of the "labeled" set such that the performance of the classifier is gradually improved. This learning theme may suffer from semantic drifting caused by errors in selecting the confident tracks. To address this issue, we propose to treat the selected frames as *related samples*—an intermediate state between labeled and unlabeled instead of labeled as in the traditional approach. To evaluate the performance, we construct a new dataset, which includes 3000 static images and 2700 face tracks of 30 celebrities. Comprehensive evaluations on this dataset and a public video dataset indicate significant improvement of our approach over established baseline methods.**

***Index Terms*—Adaptive learning, celebrity identification, related samples, semi-supervised learning, video context.**

I. Introduction

With explosive development of social network and video sharing websites, an efficient and accurate way to index and organize images and videos according to the identities of the involved persons becomes heavily demanded. Consequently, automatic character identification [1], [2], [3], which detects character faces in photos or movies and associates them with corresponding names, has attracted lots of attention in computer vision. Among many applications of character identification, celebrity-related tasks draw the most attention due to the common interest of people in celebrities. Furthermore, celebrity identification has been considered as a crucial step for image/video semantic analysis [3], [4], [5] with growing research enthusiasm in multi-media technologies.

To this end, researchers have proposed many methods for celebrity identification [6], [7], [8]. Nevertheless, as mentioned in [1] the problem still remains tremendously challenging due to: 1) lack of precisely labelled training data; 2) significant visual variations in terms of human pose, light, facial expression, etc.; 3) low resolution, occlusion, nonrigid deformation, large motion blur and complex background in the realistic photographic conditions.

An intuitive way to deal with these challenges is to collect a large-scale face database with sufficient data diversity and reliable ground-truth label. However, the enormous amount of manual work required in data labeling hinders constructing such a dataset. On the other hand, the rapid development of the Internet provides easy access to a large collection of unlabeled face data. Commercial search engines, such as Google, can return a large pool of images corresponding to a certain celebrity within just several milliseconds. Large video sharing platforms, such as YouTube, receive around 100 hours of videos uploaded every minute. The massive data available online and the easy accessibility have motivated researchers to investigate how to improve the performance of traditional learning based multimedia analysis methods utilizing such a large unlabeled dataset. As a result, semi-supervised learning [9], [10], [11] has drawn plenty of research interest during the past few decades.

In this work, we propose a novel way of utilizing video context to boost the recognition accuracy for celebrity identification with limited labeled training images. Compared with face images returned by search engines, faces in videos are captured in an unconstrained way and present more variations in pose, illumination and so forth. Moreover, although noisy, videos are usually accompanied by reliable context information that can be used for de-noising. In this paper, we extract face tracks from downloaded videos and build the celebrity identification framework with a simple but effective assumption, i.e. faces from the same face track belong to the same celebrity. More specifically, our system learns a weak classifier from a few labeled static images. The learned classifier is then used to predict the labels and confidence scores of all the frames within each video track. The frames are ranked with regard to the confidence scores and the track possessing the frame with highest confidence score is chosen as the *confident* track. The video constraint enables the

Manuscript received July 30, 2013; revised December 11, 2013 and March 21, 2014; accepted March 28, 2014. Date of publication April 10, 2014; date of current version July 15, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheng-Wei Chen. *(Corresponding author: G. Gao.)*

C. Xiong and T.-K. Kim are with the Department of Electrical and Electronic Engineering, Imperial College, South Kensington Campus, London SW7 2AZ, U.K. (e-mail: chao.xiong10@imperial.ac.uk; tk.kim@imperial.ac.uk).

G. Gao is with the School of Software, Beijing Institute of Technology, Beijing 100089, China (e-mail: guangyu.ryan@gmail.com).

Z. Zha is with Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China (e-mail: zhazj@iim.ac.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, 119077, Singapore (e-mail: eleyans@nus.edu.sg).

H. Ma is with Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: mhd@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2316475

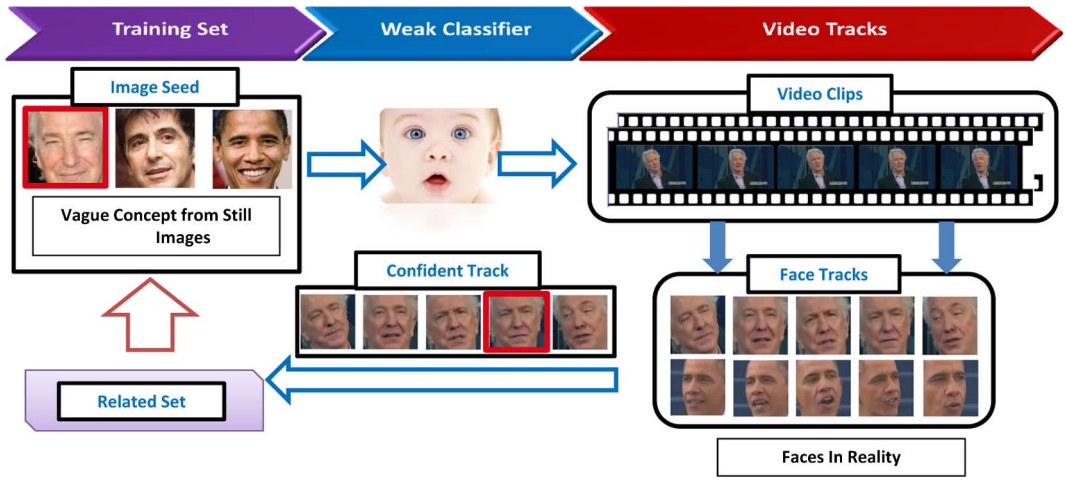


Fig. 1. Illustration of the proposed adaptive learning framework. The initial classifier is trained on a small set of static images (image seeds), and then used to label the frames within each video track. If a certain frame is assigned with a confident label, all the frames within the same track are promoted into the *related set* and utilized to update the classifier in the next iteration such that the classifier gradually evolves.

propagation of predication labels across the frames of the confident track, which is then promoted into the *related set*, as illustrated in Fig. 1. The update of the classifier is realized under the supervision of *related samples* in the *related set*. This select-update process is iterated for multiple times such that the classifier evolves with improved discriminative capacity gradually. The proposed learning theme has certain analogy to some recent biological studies of the cognitive process of human brains. According to Adaptive Resonance Theory (ART) [12], human brains form the resonant states depicting the links between visual inputs and semantics in the initial learning stage and search for the good enough matches to enhance the understanding of objects or people gradually in an adaptive learning manner.

The proposed method shares certain similarity to self-training [13], [14] since both of them adopt a mechanism of iteratively selecting samples from the unlabeled set to improve the performance. The difference lies in that our approach introduces the video context constraint into the selection process, such that the positive samples that cannot be recognized confidently may still be promoted. Self-training often suffers from the well-known semantic drifting [15]. It occurs when the size of the labeled set is too small to constrain the learning process. More specifically, the errors in selecting the *best* samples may accumulate, and consequently newly added examples tend to stray away from the original concept. Existing solutions to semantic drifting mainly focus on improving the accuracy in the selecting process, among which co-training and active learning are two major research directions. This paper, on the contrary, explores from a different perspective. Instead of struggling to select the correct samples, we aim to design a classifier robust to the selection errors by treating selected samples as *related* rather than "labeled". More specifically, we decrease the influence of the selected samples, or *related samples* as termed in this work, to guarantee that their influence is weaker than labeled samples. Furthermore, the influence of a specific *related sample* is re-weighted based on the corresponding confidence score, so that discriminative samples are emphasized while noisy and none-discriminative samples are suppressed at the same time.

II. Related Work

Celebrity identification is a specific application of face recognition. Previous works on celebrity identification can be generally categorized into two groups: a) face recognition considering correspondence between face and text information; b) face recognition utilizing a large manually labeled image or video training set.

In the first group, the textual information is used to provide extra constraint in the learning process. An early work of Satoh *et al.* [3] introduced a system to associate names located in the sound track with faces. Berg *et al.* [6] built up a large dataset by crawling news images and corresponding captions from Yahoo! News. Everingham *et al.* [2] explored textual information in scripts and subtitles and matched it with faces detected in TV episodes. However, the main disadvantage in the studies of the first group is the heavy dependence on associated textual information. In most cases, nevertheless, the assumption that textual information is available does not always hold, and errors may occur in the given text description.

The other group aims at learning a discriminative model based on a manually labeled dataset. For example, Tapaswi *et al.* [16] presented a probabilistic method for identifying characters in TV series or movies, and the face and speaker models were trained on several episodes with manual labeling. In the work of Liu *et al.* [17], a multi-cue approach combining facial features and speaker voice models was proposed for major cast detection. However, the performance of supervised learning methods mentioned above was usually constrained by the insufficiency of labeled training samples. Much research interest has been drawn by scenarios where only a limited number of labeled training samples are available, which are much more common in reality.

To overcome this scarcity in the training set, Semi-Supervised Learning (SSL) based methods [9], [10], [11] are proposed in many studies based on the assumption that unlabeled data contain the information of underlying distribution and thus can facilitate the learning process. The explosive development of

video sharing websites, such as YouTube, provides easy access to such a large unconstrained and unlabeled training set. Plentiful studies have been conducted using video data in multiple active fields of computer vision, including object detection [18], [19], object classification [20], person identification [21], action recognition [22], and attribute learning [23].

Among various SSL methods, one of the classic is the bootstrapping based method, also known as self-training. For instance, Cherniavsky *et al.* [24] trained a classifier on a set of static images and used it to recognize attributes in videos. Chen *et al.* [22] addressed the action recognition task by learning generic body motion from unconstrained videos. In their example-based strategy, the most confident pose is located in a nearest-neighbor manner and then added into the training set. Kuettel *et al.* in [25] proposed a segmentation framework on the ImageNet dataset by recursively exploiting images segmented so far to guide the segmentation of new images in a bootstrapping manner. Choi *et al.* [23] proposed to expand the visual coverage of training sets by learning from confident attributes of unlabeled samples. It was also claimed that even though some attributes were selected from other categories, they could lead to improvement in category recognition accuracy.

A typical issue in the self-training methods is caused by the error in labeling confident samples in each iteration - early errors will accumulate by including more and more false positive samples, causing semantic drifting as mentioned in [15]. Most researchers solve this problem by trying to increase the labeling accuracy in selection. Conventional approaches include co-training [26] and active learning [27]. Active learning iteratively queries the supervision of the users on the least certain samples. Li and Guo [28] proposed an adaptive active learning method by introducing a combined uncertainty measurement. They selected the most uncertain samples to query user's supervision. These selected samples are added into the training set and used to re-train the classifier. Co-training or multi-view learning, on the other hand, learns a classifier on several independent feature sets or views of data [27] or learns several different classifiers from the same dataset [29]. Saffari *et al.* [30] proposed a multi-class multi-view learning algorithm, which utilized the posterior estimation of one view as a prior for classification in other views. In [31], Minh *et al.* introduced RKHS of vector-valued functions into manifold regularization and multi-view learning, and achieved the state-of-the-art performance.

Incremental learning or online learning [32], [33] also includes a mechanism of iteratively updating the classifier. A common assumption is that the training samples with labels are given as in a streaming manner, i.e. not all the training samples are presented at the same time. Incremental learning cannot select the confident unlabeled data as in self-training and its performance is quite sensitive to the label noises. In this work, we focus on learning a robust classifier with noisy selected samples. Thus, incremental learning is out of scope in this work.

We propose an adaptive learning approach for celebrity identification by incorporating the video context information. Moreover, we introduce the concept of *related sample* to address the problem of semantic drifting. Instead of struggling to prevent the error in labeling unknown samples, we aim to obtain a classifier that is robust to selection errors such that the performance can be improved steadily.

III. Overview of Adaptive Learning

Adaptive Resonance Theory (ART) [12] is a cognitive and neural theory to describe how the brain learns to categorize in an adaptive manner. According to ART, human brain initializes the resonant states, which links the visual inputs to semantics, via "supervised learning" and then tries to find "good enough" matches for the concept in everyday life. These matches are then used for updating the resonant states in the learning process.

According to ART, the baby may learns in a two-stage manner–initial learning and adaptive learning.

- Initial Learning. A new born baby has not much knowledge, i.e. resonant states, of recognizing a certain object or person. Parents, acting as supervisors, show the baby the links between words (labels) and visual information and provide some initial labeled samples.
- Adaptive Learning. The baby observes the world by himself/herself. When a certain status of a person matches with the initial pictures in the brain (good match), the baby connects all the visual information of this person with the existing knowledge to update.

Sharing the similar spirit, our framework includes a two-stage learning mechanism on a training dataset consisting of: a) labeled images for Initial Learning and b) unlabeled noisy data from the Internet for Adaptive Learning. The images are retrieved from Google image using the name of each celebrity as the query word and then manually labeled. For collecting the noisy data, we download video clips from YouTube with tags relevant to each celebrity. Faces in the static images online are usually taken under similar conditions, e.g., similar pose, facial expression and illumination. However, faces in the videos present more variations and thus provide more diverse training samples for Adaptive Learning. Note that the collected videos are noisy due to: 1) the videos may not be relevant to the celebrity and wrongly selected due to the tagging errors of the users and 2) each video may contain several individuals. Thus such videos are treated as unlabeled data and fed into the classifier without using the ground-truth identification during training.

In this paper, we extract multiple face tracks from the collected videos and exploit the video context information within the face tracks. We introduce the *video* constraint into the adaptive learning process, i.e., faces from the same track belong to the same identity. The video constraint has a natural connection with the "baby learning" process, as mentioned in the above section. The visual perception of the baby is continuous and the baby is able to tell the correspondence between the consecutive frames, i.e whether these frames share the same identity. Namely, the baby organizes the visual perceptions in the real world as tracks of consecutive frames that belong to the same identity. Our proposed video constraint possesses a similar spirit.

Before introducing the details of our methods, we define some notations here for formal description. Suppose we are given

in total n training samples of N individuals, which include l labeled samples and u unlabeled samples (video tracks), i.e., $n = l + u$. We denote the initial labeled image set as $\mathcal{L}_o = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, where y_i represents the label for the sample x_i. The unlabeled video set consists of K face tracks $\{\mathcal{T}_i | i \in \{1, 2, \ldots, K\}\}$ with $K \leq u$, and is denoted as $\mathcal{U} = \{(x_{l+1}, \ldots, x_{l+u})\}$. Here $\{x_i, i = l+1, \ldots, n\}$ are extracted frames from the face tracks.

The most straightforward way of utilizing the unlabeled samples is to treat the most confident unlabeled track $\mathcal{T}_i$ as labeled based on the corresponding confidence score. Here the confidence score can be computed based on the classifier learned from a few labeled samples. These tracks are termed as confident tracks, which correspond to the "good enough match" in baby learning process [12]. All the frames within are then assigned with the same label as the most confident frame and promoted into the *related set* denoted as $\mathcal{L}_r$ (details in Section IV). Afterwards, the classifier is re-trained with the current "labeled set", the union of initial labeled set and discovered related set, $\mathcal{L} = \mathcal{L}_o \cup \mathcal{L}_r$. The updated classifier then predicts the labels of all the remaining frames in the video set. To identify multiple celebrities, the classifier is trained in a one-vs-all manner. More specifically, we train N binary classifiers, each of which is learned by taking one class of samples as positive and the remaining $N-1$ classes of samples as negative. The most confident tracks are then selected per class in each iteration. This is aimed to avoid the dominance of a certain class in the track selection and balance the response magnitude of all the classifiers. The confidence score of each frame belonging to class j is computed via a soft-max function $g_j(\cdot)$ on the response of each classifier:

$$g_j(x_i) = \frac{\exp\{f_j(x_i)/\eta\}}{\sum_k \exp\{f_k(x_i)/\eta\}}, \tag{1}$$

where $f_j(\cdot)$ denotes the binary classifier for the class j and η is a trade-off parameter for approximating the max function. Large η renders almost the same scores for different inputs, while small η enlarges the gaps among the output confidence scores.

We compute the confidence scores of all the frames within each face track. The maximum of these confidence scores within each track is denoted as MaxF, and the minimum is denoted as MinF. Different face tracks are ranked in terms of their MaxF scores and only the top S tracks are selected as candidates for the following selection. The candidate tracks are then ranked in terms of their MinF scores, and the track with the largest MinF score is selected as the confident track. This selection process is graphically illustrated in Fig. 2. Using this mechanism, we aim to choose the track in which a certain frame is recognized as the "best match", and the rest frames are considered to be "good enough" matches. For better understanding of this proposed mechanism, consider an extreme case where there are a large number of candidate tracks. For this case, we actually selects the most confident tracks by the averaged confidence scores of all the tracks. However, the selection results for this setting are possibly the tracks with minor between-frame variation. This may limit the generalization performance of the learned classifier. On the other hand, if S is too small, for example $S = 1$, it is quite likely to include false tracks especially when the initial classifier is trained on a small labeled set. Considering the total number of video tracks (around 2700) in our experiments, we empirically set $S = 5$ throughout the experiments. This small value of S may achieve a good trade-off between the diversity of the chosen tracks and the selection accuracy. The framework of adaptive learning based on such a selection strategy is described in Algorithm 1.

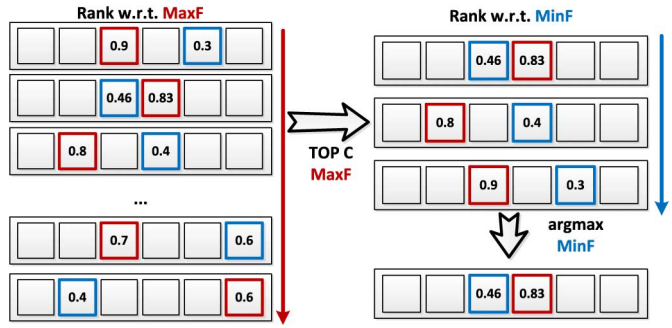


Fig. 2. Illustration of confident tracks selection mechanism. Each large block represents a face track. The small red block refers to the most confident track and the blue block refers to the least confident track. Their corresponding confidence scores are shown inside. The first selection step (left) is based on MaxF and the second step (right) is based on MinF.

Algorithm 1 Framework of Adaptive Learning.

Input:

Initial Labeled Set $\mathcal{L}_o$, Related Set $\mathcal{L}_r = \emptyset$, Unlabeled Set $\mathcal{U}$, number of classes N, maximal iteration number N_{iter}, and S for TOP-S setting.

Output: Final Classifier $F = \{f_1 \ldots, f_N\}$

1: **for** $i = 1 : N_{iter}$ **do**

2: $\mathcal{L} \leftarrow \mathcal{L}_o \cup \mathcal{L}_r$

3: Train classifier $F^{(i)} = \{f_1^{(i)}, \ldots, f_N^{(i)}\}$ on $\mathcal{L} \cup \mathcal{U}$

4: Compute $g_k(x_j), \forall x_j \in \mathcal{U}, k = \{1, \ldots, N\}$

5: **for** $k = 1 : N$ **do**

6: Compute $MaxF$ for each track.

7: Choose top S tracks as candidates according to $MaxF$.

8: Select track $\mathcal{T}_p$ with $\max MinF$ from S candidates

9: Set labels for $x_j \in \mathcal{T}_p$ as k

10: $\mathcal{L}_r \leftarrow \mathcal{L}_r \cup \{x_j \in \mathcal{T}_p\}$

11: $\mathcal{U} \leftarrow \mathcal{U} \backslash \{x_j \in \mathcal{T}_p\}$

12: **end for**

13: **end**

In general, Adaptive Learning is more robust to various changes in terms of pose, facial expression and so forth. Unlike traditional semi-supervised learning, confident samples in Adaptive Learning obtain much higher influence than the remaining unlabeled samples in the next iteration of training. With the introduced video constraint, the labels are propagated from confident frames to those frames that are difficult to label based on information from the limited initial image seeds. The promoted unconfident frames usually contain faces with more variations compared with the initial labeled samples. As a result, the classifier is trained with enriched "labeled data" with high diversity, and thus gains improvement on its generalization performance.

IV. Adaptive Learning with Related Samples

The aforementioned straightforward adaptive approach simply treats *related samples* exactly the same as labeled samples in $\mathcal{L}_o$. Such an approach only works in the ideal case where no errors occur in selecting the confident tracks. However, selection errors are generally inevitable for the following two reasons: 1) poor discriminative capability of the learned classifier in the initial learning stage where the classifier is trained only with a small number of labeled images; 2) high similarity between different persons in certain frames. The errors in the selection process will cause semantic drifting [15] and degrade the performance of the classifier. To address this problem, we introduce the concept of *related samples*, which is a comprise between *labeled* and *unlabeled* samples. Selected related samples are given higher weights than the remaining unlabeled samples but lower weights than initial labeled samples in training the classifier. As a result, the initial accurately labeled data still contribute most to the learning process such that the undesired semantic drifting effect brought by promoting *related samples* is alleviated in a controlled manner. In the following subsections, we briefly review the LapSVM for semi-supervised learning, and then introduce our proposed related LapSVM, which integrates the adaptive learning and related samples together.

A. Review of LapSVM

We formulate the aforementioned ideas under the generalized manifold learning framework. In particular, we adopt Laplacian SVM (LapSVM), introduced by Belkin *et al.* [11], as a concrete classifier learning method in this work. In this subsection, we first give a brief review of LapSVM.

LapSVM is a graph-based semi-supervised learning method. A sample affinity graph is denoted as $\mathcal{G} = \{V, E\}$, where V represents the set of nodes (data samples) and E refers to edges whose weights specify pair-wise similarity defined as follows

$$W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2), \tag{2}$$

where σ is a parameter controlling the similarity based on sample Euclidean distance and is determined via cross-validation in this work.

In LapSVM [11], classifier f is learned by minimizing the following objective function:

$$\mathcal{J}(f) = \sum_{i=1}^{l} \max(1 - y_i f(x_i), 0) + \gamma_A \|f\|_A^2 + \gamma_I \|f\|_I^2, \tag{3}$$

where $\|f\|_A^2$ represents the regularization in corresponding Reproducing Kernel Hilbert Space (RKHS) to avoid over-fitting. $\|f\|_I^2$ embodies the smoothness assumption on the underlying manifold, i.e. samples with high similarity have similar classifier responses. Here, we adopt a graph-based manifold regularizer as $\|f\|_I^2 = \sum_{i,j} (f(x_i) - f(x_j))^2 W_{ij}$.

By defining the classifier in the RKHS according to the representer theorem [34], we have the following classifier representation:

$$f(\cdot) = \sum_{i=1}^{l+u} \alpha_i k(x_i, \cdot), \tag{4}$$

where $k(x_i, \cdot)$ is a kernel function in RKHS. In this work, we adopt linear kernel trading-off the performance and computational complexity, i.e., $k(x_i, x_j) = x_i x_j$.

By substituting Eqn. (4) back into Eqn. (3), the objective function is equivalently rewritten as

$$\mathcal{J}(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \max(1 - y_i f(x_i), 0) + \gamma_A \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} + \gamma_I \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{L} \boldsymbol{K} \boldsymbol{\alpha} \tag{5}$$

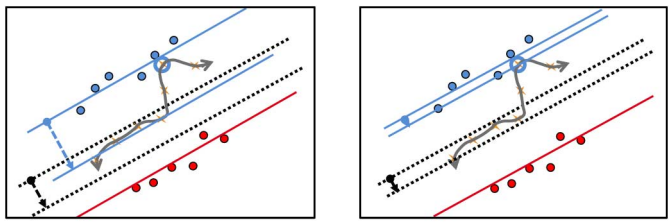
where $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}^T$, and $\boldsymbol{K}$ is the n by n gram matrix over labeled and unlabeled sample points. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the laplacian matrix on the adjacency graph $\mathcal{G}$, where $\boldsymbol{D}$ is diagonal matrix with $d_{ii} = \sum_j w_{ij}$ and $\boldsymbol{W}$ is the weight matrix defined in Eqn.(2).

LapSVM can be directly applied in our adaptive learning framework. However, as pointed out before, the cumulative error in labeling the unlabeled data may cause the problem of semantic drifting. In the following subsection, we introduce the proposed related LapSVM to solve the problem.

B. Related LapSVM

Intuitively, to solve the problem of incorrect sample selection, the influence of selected samples should be more significant than the remaining unlabeled samples, but not greater than initial original labeled samples. Referring to the LapSVM [11], labeled data are prone to be the support vectors, or in other words, lying on the margin such that $y_i(w^T x_i + b) = 1$, while there is no such constraint on unlabeled data. Selected frames, however, should lie between the decision boundary (uncertain unlabeled data) and the margin (labeled data). By considering the hard constraint in the video, frames from the same track should be put on the same half-space w.r.t. the classifier decision boundary, as shown in Fig. 3. These selected samples are treated as *related samples*, lying between the labeled and unlabeled samples.

We propose the Related LapSVM to incorporate the concept of *related sample* into LapSVM. Formally, via introducing a weight ρ for the related samples in deciding the classifier



(a) (b)

Fig. 3. Illustration on naive Adaptive Learning and Related LapSVM. Blue and red dots represent labeled samples for positive and negative class, respectively. Yellow stars represent face frames in a face track (gray curve). A certain frame (star in blue circle) is recognized as the most confident sample with a positive predicted label. Block (a) shows the change of margin (blue and red line) and decision boundary (black dashed line), as indicated by the colored arrows, for naive Adaptive Learning. Block (b) shows the change after including the concept of *related sample*. For naive adaptive learning, the margin is completely determined by selected samples, i.e., the initial labeled images are unable to constrain the learning process. However, for Related LapSVM, the influence of *related samples* do not overtake the original labeled set and the margin is retained as desired.

boundary, the objective function of LapSVM in Eqn. (5) is changed into:

$$\begin{aligned} &\mathcal{J}(\boldsymbol{\varepsilon},\boldsymbol{\alpha}) = \sum\nolimits_{i=1}^{l} \varepsilon_i + \gamma_A \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha} + \gamma_I \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{L}\boldsymbol{K}\boldsymbol{\alpha} \\ \text{s.t.}\quad & y_i \left(\sum\nolimits_{j=1}^{l+u} \alpha_j k(x_j, x_i) + b \right) \geq 1 - \varepsilon_i, \forall x_i \in \mathcal{L}_o \\ & y_{\mathcal{T}}^i \left(\sum\nolimits_{j=1}^{l+u} \alpha_j k(x_j, x_i) + b \right) \geq (\rho \cdot \mathcal{C}_{\mathcal{T}}^i) - \varepsilon_i, \forall x_i \in \mathcal{L}_r \\ & \varepsilon_i \geq 0, \forall x_i \in \mathcal{L}, 0 \leq \rho \leq 1, \end{aligned} \tag{6}$$

where ε_i is the slack variable for x_i. The predicted label $y_{\mathcal{T}}^i$ and confidence score $\mathcal{C}_{\mathcal{T}}^i$ for the most confident frame in track $\mathcal{T}_i$ are defined as follows:

$$\begin{aligned} \mathcal{C}_{\mathcal{T}}^i &= \max_{x_j \in \mathcal{T}_i} g(x_j), \\ y_{\mathcal{T}}^i &= \operatorname{sgn}(f(\arg\max_{x_j \in \mathcal{T}_i} g(x_j))), \end{aligned} \tag{7}$$

where $g(\cdot)$ is the softmax function for calculating the confidence score. With Eqn. (7), each face track is tagged with the same label as the most confident sample within.

As shown in Eqn. (6), each related sample $x_i \in \mathcal{L}_r$ is placed on a hyperplane with a distance $\rho \cdot \mathcal{C}_{\mathcal{T}}^i$ to the decision boundary. The farther the hyperplane lies away from the decision boundary, the greater influence the related samples lying on it will have in defining the decision boundary. This is based on the assumption that the track with the sample of a higher confidence score has a higher probability to be the correct track, and thus should have a stronger constraint in the training phase. The constraint in Eqn. (6) guarantees that the influence of a certain related sample is proportional to the corresponding confidence score. Also, a slack variable is imposed for each related sample, similar to the soft-margin concept in traditional SVM. ρ is a parameter in the range $[0, 1]$ to control the upper bound of the margin for related samples. A larger ρ indicates a stronger constraint on *related samples*. When ρ is set to 0, we only require all the frames within the same track to lie on the same half-space of the decision boundary.

Following the similar optimization method in [11], the problem in Eqn. (6) can be written in Lagrange form, as shown in (8) at the bottom of the page.

According to the KKT conditions, we set the derivatives of L_g in terms of b and ε_i as zeros, which yields

$$\begin{aligned} &\frac{\partial L_g}{\partial b} = 0 \Rightarrow \sum_{i, x_i \in \mathcal{L}_o} \beta_i y_i + \sum_{i, x_i \in \mathcal{L}_r} \beta_i y_{\mathcal{T}}^i = 0, \\ &\frac{\partial L_g}{\partial \varepsilon_i} = 0 \Rightarrow 1 - \beta_i - \lambda_i = 0 \Rightarrow 0 \leq \beta_i \leq 1. \end{aligned} \tag{9}$$

By substituting Eqn. (9) into Eqn. (8) and canceling b, λ, ε, the lagrangian function becomes

$$\begin{aligned} L_g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = {} & \gamma_A \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha} + \gamma_I \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{L}\boldsymbol{K}\boldsymbol{\alpha} \\ & - \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{J}_L^T \boldsymbol{Y}\boldsymbol{\beta} + \sum_{\forall i, x_i \in \mathcal{L}_o} \beta_i + \sum_{\forall i, x_i \in \mathcal{L}_r} (\rho \cdot C_{\mathcal{T}}^i)\beta_i \\ \text{s.t.}\quad & 0 \leq \beta_i \leq 1, \forall x_i \in \mathcal{L}_o \cup \mathcal{L}_r. \end{aligned} \tag{10}$$

Here $\boldsymbol{Y}$ is a diagonal labeled matrix, whose non-zero entries are set as label y_i for samples in $\mathcal{L}_o$ or predicted label $y_{\mathcal{T}}^i$ for samples in $\mathcal{L}_r$; we also define $\boldsymbol{J}_L = [\boldsymbol{I} \quad \boldsymbol{0}]$ where $\boldsymbol{I}$ is an identity matrix with a size equal to the cardinality of set $\|\mathcal{L}\|$.

Applying the KKT conditions again, we represent α by β:

$$\frac{\partial L_g}{\partial \boldsymbol{\alpha}} = 0 \rightarrow \boldsymbol{\alpha} = (2\gamma_A \boldsymbol{I} + 2\gamma_I \boldsymbol{L}\boldsymbol{K})^{-1} \boldsymbol{J}_L^T \boldsymbol{Y}\boldsymbol{\beta}, \tag{11}$$

and $\boldsymbol{K}$ is invertible since it is positive semi-definite.

Finally, the corresponding dual form of Eqn. (6) can be rewritten as follows

$$\begin{aligned} \max_{\beta}\quad & \sum_{\forall i, x_i \in \mathcal{L}_o} \beta_i + \sum_{\forall i, x_i \in \mathcal{L}_r} (\rho \cdot C_{\mathcal{T}}^i)\beta_i - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{Q}\boldsymbol{\beta}, \\ \text{s.t.}\quad & \sum_{\forall i, x_i \in \mathcal{L}_o} \beta_i y_i + \sum_{\forall i, x_i \in \mathcal{L}_r} \beta_i y_{\mathcal{T}}^i = 0, \\ & 0 \leq \beta_i \leq 1, \end{aligned} \tag{12}$$

$$\begin{aligned} Lg(\boldsymbol{\alpha}, \boldsymbol{\varepsilon}, b, \boldsymbol{\beta}, \lambda) = {} & \sum_{i=1}^{l} \varepsilon_i + \gamma_A \alpha^T \boldsymbol{K}\boldsymbol{\alpha} + \gamma_I \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{L}\boldsymbol{K}\boldsymbol{\alpha} \\ & - \sum_{\forall i, x_i \in \mathcal{L}_o} \beta_i \left(y_i \left(\sum_{j=1}^{l+u} \alpha_j k(x_j, x_i) + b \right) - 1 + \varepsilon_i \right) - \sum_{i=1}^{l} \lambda_i \varepsilon_i \\ & - \sum_{\forall i, x_i \in \mathcal{L}_r} \beta_i \left(y_{\mathcal{T}}^i \left(\sum_{j=1}^{l+u} \alpha_j k(x_j, x_i) + b \right) - \rho \cdot C_{\mathcal{T}}^i + \varepsilon_i \right). \end{aligned} \tag{8}$$

where

$$\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{J}_L\boldsymbol{K}(2\gamma_A\boldsymbol{I} + 2\gamma_I\boldsymbol{L}\boldsymbol{K})^{-1}\boldsymbol{J}_L^T\boldsymbol{Y}. \tag{13}$$

Eqn. (12) is a standard QP problem. The optimal solution can be derived utilizing traditional off-the-shelf SVM QP solvers, and we use SPM:QPC solver[1] in this work.

C. Classification Error Bound of Related LapSVM

Here we provide a theoretical classification error bound for the proposed related LapSVM, via comparing with the established error bound of standard LapSVM. Experimental performance evaluation for the related LapSVM is deferred to the section of experiments.

Given a data distribution $\mathcal{D}$ and classifier function class $\mathcal{F}$, the classification error of LapSVM is bounded by the summation of the empirical error, function complexity and data complexity, as formally stated in the following lemma [35].

Lemma 1. ([35]): Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from an input space $\mathcal{X}$ to $[0,1]$. Let $\{x_i\}_{i=1}^{l}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at leat $1-\delta$ over random draws of samples of size l, every $f \in \mathcal{F}$ satisfies

$$E_{\mathcal{D}}[f(x)] \leq \hat{E}[f(x)] + R_l(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2l}}, \tag{14}$$

where $\hat{E}[f(x)]$ is the empirical error averaged on the l examples and $R_l(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$.

By utilizing the error bound of SVM [36], $\hat{E}[f(z)] \leq O(\|\xi\|_2^2 \log^2 l)$, we can further bound the error of LapSVM in terms of the slack variable ε_i as follows,

$$E_D[f(x)] \leq O\left(\sum_i \varepsilon_i^2 \log^2 l\right) + R_l(\mathcal{F}) + \sqrt{\frac{ln(2/\delta)}{2l}},$$

where ε_i is the slack variable for sample x_i in the labeled or related sample set. The proposed related LapSVM reduces the classification error bound over LapSVM via properly re-weighting the slack variable for the unconfident/noisy samples. Specifically, consider the case where a sample x_j is selected as a confident sample but labeled incorrectly. For x_j, training the classifier actually minimizes an incorrect slack variable ε_j, and maximizes the correct slack variable $\hat{\varepsilon}_j = 1 - \varepsilon_j$, due to its opposite label. $\hat{\varepsilon}_j$ is maximized within the range of $[0,2]$. Thus, the error bound is increased to

$$E_D[f(z)] \leq O\left(\left(\sum\nolimits_{i\neq j} \varepsilon_i^2 + \hat{\varepsilon}_j^2\right)\log^2 l\right) + R_l(\mathcal{F}) + \sqrt{\frac{ln(2/\delta)}{2l}}.$$

In contrast, related LapSVM reduces the feasible range of $\hat{\varepsilon}_j$ to $[0,2] - \rho \cdot C_T^j]$. Consequently, the value of $\hat{\varepsilon}_j$ is decreased, and we have a lower error bound for related LapSVM than standard LapSVM:

$$E_D[f_{\mathrm{re-LapSVM}}(x)] \leq E_D[f_{\mathrm{LapSVM}}(x)]. \tag{15}$$

The above analysis can be generalized to the case where more unlabeled samples are labeled incorrectly. Thus we can conclude that the related LapSVM reduces error bound via handling the incorrectly labeled samples better.

V. Experiments

We conduct extensive experiments to evaluate the effectiveness of the proposed adaptive learning method for celebrity identification. This section is organized as follows. Subsection V-A introduces the details of construction of the used database. We demonstrate the experimental settings in details in subsection V-B. Subsection V-C shows a naive approach of including the video constraint in building the sample affinity graph and demonstrates that video context can improve the performance with a limited degree. Subsection V-D and V-E show the effectiveness of related samples in both supervised and semi-supervised learning scenarios. The average precision is reported on both image and video testing set. Subsection V-F illustrates the performance curve of the proposed method along with learning iterations. We also include in the last subsection experiments of related samples on a public database - YouTube Celebrities Database.

A. Database Construction

Since there are rare databases with sufficient image and video samples for celebrity identification, in this work, we construct a database for benchmarking different methods for this task. The collection of image and video data is described as follows.

1) Image Data: We select 30 celebrities who are well-known within their fields so that sufficient corresponding video data can be crawled. For each individual, we retrieve about 100 clear images from Google Image using the names of celebrities as query. We manually label all the images and mark the locations of eyes. All faces are then normalized via a standard affine transformation. There is not any strict constraint in photography conditions - different poses, facial expressions and illumination conditions are all allowed. 15 images are randomly sampled to form the training image set, while the remaining are used as the testing set. We report the average precision (AP) from 5 z-testing splits. The list of celebrities chosen in the database is given in Table I.

2) Video Data: Querying by the names of the celebrities, a video corpus consisting of about 300 video clips is downloaded from video sharing websites, e.g., YouTube. Note that for the following experiments, we assume that the videos are unlabeled for the following reasons: a) the keyword searching results are not reliable, and videos are not necessarily related with the celebrities; b) there may also be other individuals other than the celebrities of interest in the returned videos.

In this paper, only the detected face tracks are considered in the iterative adaptive learning process. Thus based on the video

[1]http://sigpromu.org/quadprog/

TABLE I
CELEBRITIES INCLUDED. WE CHOOSE PEOPLE WITH DIFFERENT OCCUPATIONS AS LISTED ABOVE. FOR DIFFERENT OCCUPATIONS, VIDEO DATA ARE COLLECTED FROM DIFFERENT VIDEO SOURCES CORRESPONDINGLY

Occupation	Name	Gender	Video Source
Politician	Barack Obama	M	Speech News Report
	Yingjiu Ma	M	
	Al Sharpton	M	
Western Actor	Adam Sandler	M	Movies Interviews
	Alexander Skargard	M	
	Alan Alda	M	
	Anthony Hopkins	M	
	Alan Rickman	M	
	Alan Tricke	M	
	Amy Poehler	F	
	Alicia Silverstone	F	
Asian Actor	Chao Deng	M	Movies Interviews
	Baoqiang Wang	M	
	Zidan Zhen	M	
	Benshan Zhao	M	
	Bingbing Fan	F	
	Wei Tang	F	
	Yuanyuan Gao	F	
Singer	Dehua Liu	M	Music Albums Concerts
	Katty Perry	F	
	Wenwei Mo	F	
	Xiaochun Chen	M	
	Yanzi Sun	F	
Hoster &Anchor	Anderson Cooper	M	News Report Talk Show TV Programs
	Fujian Bi	M	
	Lan Yang	F	
	Jing Chai	F	
CEO	Yun Ma	M	Commercial News Product Launch Video
	Bill Gates	M	
	Steve Jobs	M	

constraint, the label is transferred from confident frames to uncertain frames within the same track. Besides, by only considering the detected tracks, the volume of frames that need to be processed can be largely reduced to accelerate the learning process. To obtain reliable face tracks, a robust foreground correspondence tracker [37] is applied for each shot.

Here video shot segmentations are automatically detected with the accelerating shot boundary detection method [38]. More specifically, the Focus Region (FR) in each frame is defined, and using a skip interval of 40 frames, the method not only speeds up the detection process, but also finds more subtle transitions.

After segmenting the video into shots, the tracking process takes the results of OKAO face detection[2] as input, and generates several face tracks using the tracking algorithm in [37]. The face tracks are then further pruned via fine analysis of faces as follows:

- Duration. Short tracks with less than 30 frames are discarded, since these tracks are often introduced by false positive detections.
- Clusters. K-means clustering is applied to each track, and only those frames closest to clustering centers are chosen as corresponding representative faces.

Consequently we acquire around 2,700 video tracks in total with nearly 90 tracks per individual.

3) Feature for Face Recognition: We adopt the following three types of state-of-the-art features in face recognition: Gabor, LBP and SIFT feature. Details of the feature extraction are listed below:

- **Gabor Feature**. Gabor filter [39] has been widely used for facial feature extraction due to its capability of capturing salient visual properties, such as spatial localization, orientation selectivity as well as spatial frequency characteristics. In this paper, we adopt a common setting for extracting gabor feature: wavelet filter bank with 5 scales and 8 orientations, central frequency is set as $\sqrt{2}$, and filter window width is set as 2π.
- **Local Binary Patter Feature**. LBP captures the contrast information of the central pixel and its neighbors. The advantage of LBP lies in its robustness to illumination and pose variations. We use a variant of LBP - multi-block LBP [40]. In the feature selection, the image is firstly segmented into several blocks to keep a certain amount of geometric information. Each face image is divided into 5×4 sub-regions and then for each sub-region uniform patterns are extracted and concatenated as bins for a histogram representation.
- **SIFT Feature**. A nine-point SIFT feature is used in the experiments. Referring to the work of Everingham *et al.* [2], a generative model is adopted to locate the nine facial key-points in the detected face region, including the left and right corners of each eye, the two nostrils and the tip of the nose and the left and right corners of the mouth followed by 128-dim SIFT feature [41] extraction process.

The vectors of the above three features are normalized individually by l_2-norm and concatenated into a single vector for each image/frame.

B. Experiment Settings

In the following experiments, the initial training image set is constructed by randomly sampling 15 images per person from the labeled image data, and the rest images are used for testing. We run this sampling process for 5 times in each experiment and report the mean precision in this paper.

We consider two scenarios for experiments: 10-person and 30-person scenario. In the 10-person scenario, 10 celebrities are selected randomly from the name list in Table I and corresponding training samples are chosen as above. We perform such random selection processes for 3 times and then reported the average precision (AP). In the 30-person scenario, we use the training samples of all celebrities.

For AL, we follow the procedures in Algorithm 1 with the value of parameter S set as 5. The maximal iteration number is set as $N_{iter} = 15$, and the results for AL based approaches are the accuracy of the final learning iteration. The parameter η in Eqn. (1) is set as 0.7. In Related LapSVM defined in Eqn. (6), γ_I and γ_A are set as 10^{-2} and 1, respectively, and ρ is empirically set as 0.3.

C. Video Constraint in Graph

LapSVM [42] is a graph-based classifier and we take a baseline extension to incorporate the video context information into LapSVM framework.

The general idea is to include the video constraint when constructing the affinity matrix, which defines the similarity among

[2]http://www.omron.com/r_d/coretech/vision/okao.html

TABLE II
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT SVM BASED METHODS IN THE 10-PERSON SCENARIO

		3	5	7	10	12	15
Image	Lap+V	**50.83**	**60**	**68.33**	76.67	82.5	84.17
	SVM	39.17	48.75	58.75	75	75.83	78.75
	ST-SVM	41.67	51.25	61.25	75.42	76.25	80.83
	AL-SVM	43.34	51.25	58.75	77.92	80.84	82.09
	Re-SVM	50.42	54.17	65.42	**82.5**	**82.92**	**84.59**
Video	Lap+V	**53.09**	48.38	49.97	56.62	70.32	73.41
	SVM	30.4	33.09	45.44	62.23	66.76	70.68
	ST-SVM	29.11	34.17	44.21	64.12	66.88	72.93
	AL-SVM	46.32	50.3	45.84	55.73	63.37	75.42
	Re-SVM	49.55	**50.45**	**50.57**	**76.26**	**79.26**	**78.18**

TABLE III
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT SVM BASED METHODS IN THE 30-PERSON SCENARIO

		3	5	7	10	12	15
Image	Lap+V	**48.87**	**62.83**	**72**	**82**	**84.5**	**86.5**
	SVM	39.33	51.17	61.67	75	79.5	82.17
	ST-SVM	39.33	50.33	60.5	75.33	79	81.67
	AL-SVM	38.34	48.17	57.34	72.17	79	84.5
	Re-SVM	41.5	52.84	63	76.34	82.5	84.34
Video	Lap+V	42.13	46.48	45.93	56.61	62.39	60.82
	SVM	31.97	38.79	41.41	50.88	60.48	64.95
	ST-SVM	32.26	38.08	40.74	51.55	59.67	63.61
	AL-SVM	36.06	38.35	48.95	64.01	69.55	74.02
	Re-SVM	**49.28**	**46.67**	**49.09**	**67.03**	**71.47**	**75.42**

training instances. A naive approach is to set the similarity of frames from the same track to be 1. Nevertheless, experiments show that this setting usually results in a degradation in performance. A possible reason could be that the weight among consecutive frames becomes much larger than other entries within the weight matrix, which makes the classifier dominated by the constraints on corresponding samples other than labeled instances. Therefore, to ensure the balance of sample weights, the weight is defined as the summation of graphic similarity and video constraint. In detail, the edge between consecutive frames is defined as,

$$w_{ij} = \lambda \cdot \exp\{-(x_i - x_j)^2/2\sigma^2\} + (1-\lambda)\cdot \min\{\zeta \cdot \mu_W, 1\}, \quad (16)$$

where μ_W is the mean of matrix $\boldsymbol{W}$.

In Eqn. (16), we confine $\lambda \in [0,1]$ and $\zeta \in [1,10]$ empirically. We tune the values for λ with a step of 0.1 and ζ with a step of 1 within their corresponding ranges via cross-validation. Experiments show a small improvement over LapSVM of 1% on average.

This approach is named as Lap+V, and is taken as the baseline algorithm in the following experiments.

D. Related Sample in Supervised Learning

In this subsection, we evalute the effect of *related sample* on SVM. γ_I in Eqn. (6) is set as 0 and the classifier is defined only in terms of labeled samples $f(\cdot) = \sum_{i=1}^{l} \alpha_i K(x_i, \cdot)$. We compare the following methods: SVM, ST-SVM (self-training with SVM), AL-SVM (adaptive learning with video constraint), and Re-SVM (related SVM). Similar to Section V-C, the performance is evaluated on both image and video data in 10-person and 30-person scenarios respectively. Average precision is reported in Table II and III under varying numbers of labeled training images.

For traditional self-training, only those frames with high similarity to the initial training samples are selected to enlarge the training set. Thus, the variations in the selection samples are limited. Limited number of labeled samples may decrease AP due to the high error rate during selection, while, more labeled samples usually result in improvement for ST-SVM. However, the difference for either degradation and improvement is very small: less than 1%. Straight-forward adaptive learning (AL+SVM) demonstrates similar performance, but the range for both degradation and improvement are largely increased to around 4%.

Related SVM adjusts the margin for each sample in accordance with their confidence scores, such that we can amplify the positive influence of more confident samples while suppressing the negative influence of less confident samples. Generally, by regarding selected samples as related samples, the classifier is much more robust to selection errors. As shown in Table II and III, the improvement of Re-SVM over SVM is around 5% on image data and 12% on video data. In most cases where the number of labeled samples is small (e.g. the number is 3 or 5), the initial classifier is unreliable. Normally, around half of the selected tracks are not correctly labeled by the classifier. Related SVM can significantly degrade the impact of error tracks and provide considerable AP improvement. With sufficient labeled training samples (e.g. 12 or 15), the generalization performance of the classifier is significantly improved. The error rate in selecting tracks is low, and thus correct samples play a dominant role in training. In such a case, the improvement brought by related samples becomes less significant.

Note that there is still an considerable performance gap between Related SVM and LapSVM with video constraint (Lap+V) on the image testing dataset: 3% and 6% in 10-person and 30-person cases. A possible reason lies in the fact that both training and testing samples are static images downloaded from Google. The correlation between video data and image data is low. As a consequence, the right tracks selected in AL will result in minor improvement for testing on images, while, the incorrect tracks will degrade the performance to a certain extent. The impact of error tracks is relatively significant compared with the influence of the right tracks. However, on the video dataset, Related SVM outperforms LapSVM with a margin of 5% and 7% in both 10-person and 30-person cases. Especially, when sufficient labeled samples are fed into the training process - 10 or more, the improvement can be up to 20%.

E. Related Samples in Semi-Supervised Learning

In this subsection, we examine the effect of related samples in semi-supervised learning and take LapSVM and Transductive SVM (TSVM) as the base classifiers for Adaptive Learning.

TABLE IV
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT LAPSVM BASED METHODS IN THE 10-PERSON SCENARIO

		3	5	7	10	12	15
Image	Lap+V	50.83	60	68.33	76.67	82.5	84.17
	ST-LapSVM	48.75	56.67	66.25	77.5	78.75	82.08
	AL-LapSVM	**50.84**	46.67	73.34	81.25	82.92	87.92
	Re-LapSVM	49.17	**61.67**	**75.84**	**83.34**	**85.42**	**88.34**
Video	Lap+V	53.09	48.38	49.97	56.62	70.32	73.41
	ST-LapSVM	34.83	42.39	48.62	66.34	71.19	75.72
	AL-LapSVM	48.21	16.67	39.27	64.06	63.76	79.83
	Re-LapSVM	**53.46**	**55.28**	**77.01**	**83.13**	**83.28**	**84.36**

TABLE V
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT LAPSVM BASED METHODS IN THE 30-PERSON SCENARIO

		3	5	7	10	12	15
Image	Lap+V	48.87	62.83	72	82	84.5	86.5
	ST-LapSVM	45.33	60	72.17	82.33	84.67	86.67
	AL-LapSVM	40.17	52.34	64.17	76.67	78.34	84
	Re-LapSVM	**49**	**64.67**	**72.84**	**83.84**	**84.67**	**87.5**
Video	Lap+V	42.13	46.48	45.93	56.61	62.39	60.82
	ST-LapSVM	38.22	49.5	59.18	**64.76**	70.38	69.71
	AL-LapSVM	26.41	41.52	39.75	57.32	62.68	61.75
	Re-LapSVM	**42.29**	**50.66**	**57.16**	63.33	**71.52**	**72.9**

When building up the affinity graph in LapSVM, video constraint in Eqn. (16) is not included. Related LapSVM (Re-LapSVM) is considered as another way of incorporating video constraint into the learning process other than Lap+V in Section V-C. The video context information is utilized in the process of promoting tracks into the related set.

We investigate whether further improvement of LapSVM can be brought by Re-LapSVM over Lap+V. The results are given in Table IV and V. We observe similar results in comparisons among self-training with LapSVM (ST-LapSVM), straight-forward AL (AL-LapSVM) and AL with related samples (Re-LapSVM). Re-LapSVM outperforms both ST-LapSVM and AL-LapSVM. Re-LapSVM demonstrates a better tolerance to selection errors than the AL-LapSVM, especially for cases with 3, 5 and 7 labeled samples. More importantly, the comparison between Lap+V and Re-LapSVM demonstrates more insightful results. Re-LapSVM demonstrates a significant advantage over Lap+V. In details, the enhancement on AP is around $\mathbf{4}\%$ for 10-person image case, $\mathbf{1.5}\%$ for 30-person image case, $\mathbf{16}\%$ for 10-person video case, $\mathbf{8}\%$ for 30-person video case, respectively.

In the implementation of TSVM, we optimize TSVM following Collober *et al.* [43] with concave-convex procedure (CCCP). The objective function of TSVM is non-convex, and CCCP optimizes the problem by solving multiple quadratic programming subproblems. For each QP subproblem, we follow the similar way of incorporating related samples as in Eqn. (6) in Section IV-B. Since the optimization of TSVM is slow, we only conduct experiments in 10-person scenario. The results are demonstrated in Table VI.

Clearly, the performance of TSVM is worse than that of LapSVM, especially when labeled samples are limited. However, the comparison between TSVM and LapSVM is out of the scope of this work. Here, our focus is on whether related samples improve the performance of TSVM as well. As shown in Table VI, Re-TSVM outperforms both TSVM and AL-TSVM with an improvement of around 3%.

TABLE VI
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT TSVM BASED METHODS IN THE 10-PERSON SCENARIO

		3	5	7	10	12	15
Image	TSVM	41.67	51.67	60.42	75.83	77.08	79.58
	AL-TSVM	39.58	48.75	59.17	74.58	79.17	85
	Re-TSVM	**42.5**	**52.08**	**62.5**	**80**	**78.75**	**85**
Video	TSVM	32.01	38.82	46.91	61.87	68.11	71.94
	AL-TSVM	**39.24**	40.05	43.05	61	73.56	78.66
	Re-TSVM	36.06	**42.66**	**47.39**	**75.03**	**75.99**	**81.92**

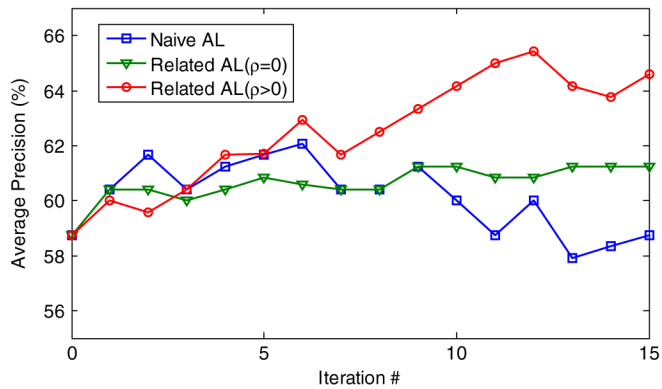


Fig. 4. Learning Curves of three approaches: Naive AL, Related AL ($\rho = 0$) and Related AL ($\rho > 0$).

F. Learning Curves of Adaptive Learning

In this subsection, we investigate the behaviors of different approaches by investigating the average precision with respect to the iteration number. In this experiment, the labeled set for testing and training is fixed for a fair comparison. The maximal iteration count is set as 15, and accuracy on testing data is reported for each iteration. Since the learning curve is similar for most simulation runs, Fig. 4 illustrates one run for LapSVM-based Adaptive Learning.

It is easy to observe that straightforward Adaptive Learning (Naive AL) shows a noisy curve since it is quite sensitive to the selection errors. If the correct track is chosen, accuracy will demonstrate an obvious increase, and the performance will drop suddenly if errors occur in the process of selection. Re-LapSVM with $\rho = 0$ shows a smooth learning curve and converges. Re-LapSVM with $\rho > 0$ shares the similar behaviors of the two approaches to some extent: the trend of AP is increasing but with minor turbulence. The parameter ρ in Eqn. (6) is an important factor controlling the relative influence compared with the labeled image samples in the learning process. Larger ρ will render the learning curve closer to straightforward AL, while smaller ρ pushes the learning curve towards related AL with $\rho = 0$. An exemplar illustration of simulation results is also presented in Fig. 5. In general, the observed results are consistent with our expectation.

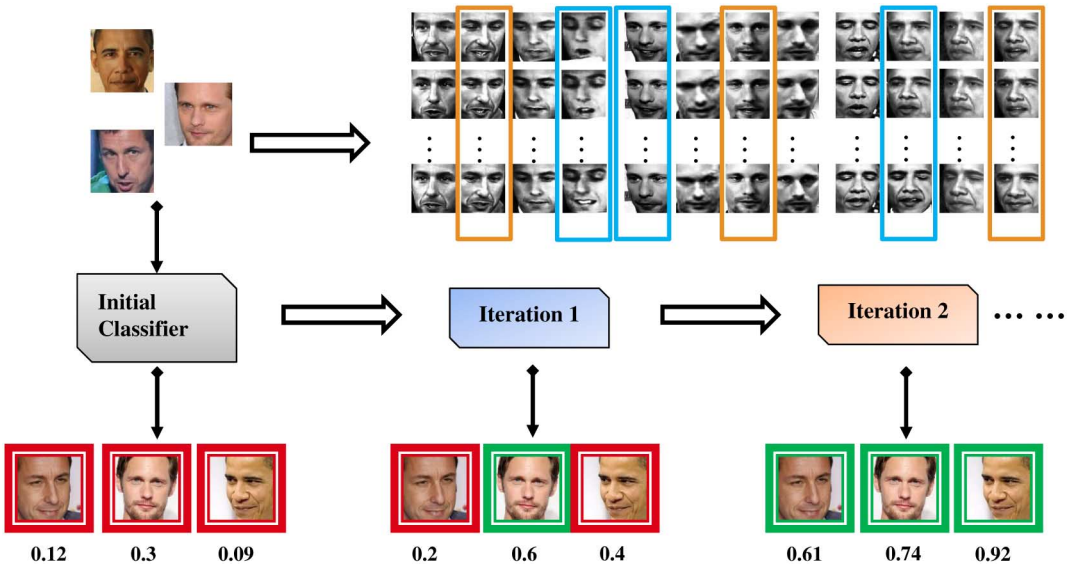


Fig. 5. Examples of iterative improvement. The upper left static images are used for training the initial classifier, and the gray image matrix represents the pool of video tracks with each column standing for a track. In Iteration 1, tracks in blue bounding box are chosen, while in Iteration 2, tracks in orange bounding box are selected. The lowermost row are examples of testing images with corresponding confidence scores shown below. Red frame indicates wrong decision and green frame indicates right decision. With more tracks selected into the training pool, the confidence score on the testing dataset is rising.

TABLE VII
COMPARISON ON THE AVERAGE PRECISION (%) OF DIFFERENT LAPSVM BASED METHODS IN YOUTUBE CELEBRITIES DATABASE

	3	5	7	10	12	15
SVM	45.73	47.3	55.05	57.66	62.25	41.33
ST-SVM	43.35	47.47	55.07	57.39	63.36	63.5
Lap+V	49.97	50.93	60.37	63.83	67.81	68.25
ST-LapSVM	51.3	52.3	61.55	63.92	68.05	68.29
AL-LapSVM	55.93	55.72	62.02	65.33	68.59	68.8
Re-LapSVM	**58**	**58.09**	**65.61**	**65.81**	**69.81**	**68.87**

G. YouTube Celebrity Dataset

We also evaluate the proposed algorithm on a public dataset–the YouTube Celebrity Dataset [44], which contains 1,910 sequences of 47 subjects. All the sequences are extracted from video clips downloaded from YouTube by evicting frames that do not contain celebrites of interest. Most of the videos are of low resolution and recorded at high compression rates. The size of frames ranges from 180×240 to 240×320 pixels.

Following the similar methods described in Section V-A, face tracks are extracted within each video sequence. Only celebrities with more than 30 tracks are included in this experiment and the final number of identities is 32. Since there is no separate image set for the initial training stage as in our approach, we randomly sample 5 tracks for each celebrity. All the frames within are then treated as initial labeled samples. This sampling process is repeated for 5 times and the corresponding averaged results are shown in Table VII. The results are similar to those observed on our own dataset and the improvement of Re-LapSVM over Lap+V is around 4% on average.

Compared with the results on our own dataset, the improvement of Related LapSVM is less significant over the baseline algorithms. The reason is that the proposed method targets at solving a common problem in real applications, namely it is difficult to collect many training images to train reliable initial classifiers. When the number of labeled training images is small, the classifiers are not reliable, and errors in selecting the confident video tracks by such weak initial classifiers are inevitable. In this case, the performance of classifiers may degrade severely due to incorporating more and more noisy or incorrect samples. Thus, the improvement brought by related LapSVM is more significant with more noisy tracks selected.

Compared with the proposed dataset, the error rate in selecting confident tracks on Youtube dataset is much lower. Thus the performance gain of Re-LapSVM is smaller on the Youtube dataset compared with on our own dataset. The reasons of lower track selection error on the Youtube dataset are two-fold: 1) The Youtube dataset only contains videos, so we train the initial classifier using the video data. Such video-domain classifiers perform more accurately in selecting the confident remaining video tracks than the initial classifiers trained from image-domain in our own dataset; 2) The face sequences (tracks) for each individual in Youtube faces dataset are usually extracted from only 2-3 videos, and the correlation/similarity among different sequences from the same video is quite high. However, the dataset built in this work contains tracks from about 10 different videos for each celebrity. Thus, our video dataset is much more diverse and difficult for track selection than the Youtube face dataset. Due to the above two reasons, the performance improvement on Youtube dataset achieved by Re-LapSVM is less significant than that on our dataset.

VI. CONCLUSION

A novel adaptive learning framework was proposed for the celebrity identification problem inspired by the concept of "Baby learning". The classifier is initially trained on labeled static images, and gradually improves by augmenting confident face tracks into the knowledge base. We also proposed a robust classifier that is robust to selection errors by assigning weak

adaptive margin for those selected samples. Extensive experiments are conducted in both supervised and semi-supervised learning setting for celebrity identification. Results on two databases show that the improvement on accuracy is significant and inspiring. Although in this work we only consider the task of celebrity identification, the proposed method is a general approach and can be easily extended to solve other problems in computer vision as well, such as object detection, object recognition and action recognition.

REFERENCES

[1] O. Arandjelovic and A. Zisserman, "Automatic face recognition for film character retrieval in feature-length films," in *Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2005, pp. 860–867.

[2] M. Everingham, J. Sivic, and A. Zisserman, "'hello! my name is… buffy' - automatic naming of characters in tv video," in *Proc. 17th Brit. Machine Vision Conf.*, 2006, pp. 889–908.

[3] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: Naming and detecting faces in news videos," *IEEE Multimedia Mag.*, vol. 6, no. 1, pp. 22 –35, Jan.-Mar. 1999.

[4] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies," *Multimedia Tools Appl.*, vol. 48, no. 2, pp. 313–337, Jun. 2010.

[5] M. Bertini, A. Del Bimbo, and W. Nunziati, "Automatic detection of player's identity in soccer videos using faces and text cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 663–666, ser. MULTIMEDIA '06.

[6] A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth, "Names and faces in the news," in *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2004, pp. II-848–II-854.

[7] M. Zhao, J. Yagnik, H. Adam, and D. Bau, "Large scale learning and recognition of faces in web videos," in *Proc. 8th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2008, pp. 1–7.

[8] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 995–1007, Aug. 2012.

[9] X. Zhu, J. Lafferty, and Z. Ghahramani, "Semi-supervised learning: From Gaussian fields to Gaussian processes," in *Proc. 2003 Int. Conf. Mach. Learn.*, 2003.

[10] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Advances Neural Inf. Process. Syst.*, 2003.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[12] S. Grossberg, "Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world," *Neural Netw.*, vol. 37, pp. 1–47, 2013.

[13] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annu. Meet. Assoc. Comput. Linguistics*, 1995, pp. 189–196.

[14] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," in *Proc. 2006 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2006.

[15] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning using attributes and comparative attributes," in *Proc. 12th Eur. Conf. Comput. Vision*, 2012.

[16] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "'Knock! Knock! Who is it?' probabilistic person identification in tv series," in *Proc.2012 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, Jun. 2012.

[17] Z. Liu and Y. Wang, "Major cast detection in video using both speaker and face information," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 89–101, Jan. 2007.

[18] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *Proc. 2013 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2013, pp. 1650–1657.

[19] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. 2012 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2012, pp. 3282–3289 .

[20] R. Yan, J. Zhang, J. Yang, and A. G. Hauptmann, "A discriminative learning framework with pairwise constraints for video object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 578–593, Apr. 2006.

[21] M. Bauml, M. Tapaswi, and R. Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data," in *Proc. 2013 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2013, pp. 3602 –3609.

[22] C.-Y. Chen and K. Grauman, "Watching unlabeled video helps learn new human actions from very few labeled snapshots," in *Proc. 2013 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2013, pp. 572–579.

[23] J. Choi, M. R. Ali, F. Larry, and S. Davis, "Adding unlabeled samples to categories by learned attributes," in *Proc. 2013 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2013, pp. 875–882.

[24] N. Cherniavsky, I. Laptev, J. Sivic, and A. Zisserman, "Semi-supervised learning of facial attributes in video," in *Proc. 2010 Eur. Conf. Comput. Vision Workshop*, 2010.

[25] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Proc. 2012 Eur. Conf. Comput. Vision Workshop*, 2012.

[26] A. Fathi, M. F. Balcan, X. Ren, and J. M. Rehg, "Combining self training and active learning for video segmentation," in *Proc. 2011 Brit. Mach. Vision Conf.*, 2011.

[27] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, New York, NY, USA, 1998, pp. 92–100, ser. COLT' 98.

[28] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Proc. 2013 IEEE Conf. Comput. Vision Pattern Recog.* , 2013, pp. 859–866.

[29] S. Goldman and Y. Zhou, "Enhancing supervised learning with unlabeled data," in *Proc. 2000 Int. Conf. Mach. Learn.*, 2000.

[30] A. Saffari, C. Leistner, M. Godec, and H. Bischof, "Robust multi-view boosting with priors," in *Proc. 2010 Eur. Conf. Comput. Vision*, 2010, vol. 6313, pp. 776–789, ser. Lecture Notes in Computer Science.

[31] H. Q. Minh, L. Bazzani, and V. Murino, "A unifying framework for vector-valued manifold regularization and multi-view learning," in *Proc. 30th Int. Conf. Mach. Learn.*, May 2013, vol. 28, no. 2, pp. 100–108.

[32] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2006, pp. 260–267.

[33] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. 10th Eur. Conf. Comput. Vision: Part I*, 2008, pp. 234–247, ser. ECCV '08.

[34] B. Scholkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *Computational Learning Theory*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2001, vol. 2111, pp. 416–426.

[35] S. Sun, "Multi-view laplacian support vector machines," in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2011, pp. 209–222.

[36] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[37] S. Wang, H. Lu, Y. F. , and Y. M. H. , "Superpixel tracking," in *Proc. 13th Int. Conf. Comput. Vision*, 2011, pp. 1323–1330.

[38] G. Gao and H. Ma, "Accelerating shot boundary detection by reducing spatial and temporal redundant information," in *Proc. 2011 IEEE Int. Conf. Multimedia Expo.*, 2011, pp. 1–6.

[39] J. G. Daugman, "Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A-Optics Image Sci. Vision*, 1985.

[40] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[41] Lowe and D. G. , "Object recognition from local scale-invariant features," in *Proc. 1999 Int. Conf. Comput. Vision*, 1999.

[42] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *J. Mach. Learn. Res.*, vol. 12, pp. 1149–1184, Mar. 2011.

[43] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Large scale transductive svms," *J. Mach. Learn. Res.*, vol. 7, pp. 1687–1712, Dec. 2006.

[44] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. 2008 IEEE Conf. Comput. Vision Pattern Recog.*, 2008, pp. 1–8.

Chao Xiong received the B.Sc. degree in engineering of telecommunication from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the M.Sc. degree in communication and signal processing from the Department of Electrical and Electronic Engineering, Imperial College, London, U.K., in 2011. He is currently a Ph.D. candidate at the Department of Electrical and Electronic Engineering, Imperial College, London, U.K.

His research interests include computer vision and pattern recognition.

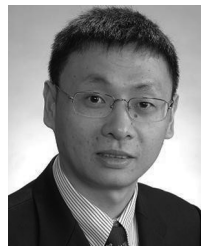
Guangyu Gao received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013, and the M.S. degree in computer science and technology from Zhengzhou University, Henan, China, in 2007. He was a government-sponsored joint-Ph.D. student at the National University of Singapore, Singapore, from July 2012 to April 2013.

He is currently an Assistant Professor at the School of Software, Beijing Institute of Technology, Beijing, China. His current research interests include applications of multimedia, computer vision, video analysis, machine learning, and big data.

Zhengjun Zha (M'08) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China.

He previously worked as a Senior Research Fellow in the School of Computing, National University of Singapore, Singapore. He is currently a Professor at the Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, China. His current research interests include user-generated content analysis, multimedia information retrieval, and social media analysis.

Dr. Zha has published a series of book chapters, journal articles, and conference papers in the areas of user-generated content analysis, multimedia information retrieval, and social media analysis, including articles that appeared in IEEE TRANSACTIONS ON MULTIMEDIA, ACM TRANSACTIONS ON MULTIMEDIA COMPUTING, COMMUNICATIONS, AND APPLICATIONS, IEEE TRANSACTION ON IMAGE PROCESSING, ACM International Conference on Multimedia (ACM Multimedia), the Conference on Computer Vision and Pattern Recognition, and the Special Interest Group on Information Retrieval. He received the Best Paper Award at ACM Multimedia 2009, the Best Demo Runner-Up award at ACM Multimedia 2012, the Best Student Paper Award at ACM Multimedia 2013, and the Best Paper Award at the International Conference on Internet Multimedia Computing and Service 2013.



Shuicheng Yan (SM'13) is currently an Associate Professor in the Department of Electrical and Computer Engineering at National University of Singapore, Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). His research areas include computer vision, multimedia, and machine learning.

Dr. Yan has authored and co-authored over 300 technical papers over a wide range of research topics, with Google Scholar citation 8100 times and H-index-40. He is an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT) and ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY (ACM TIST), and has been serving as the Guest Editor of the special issues for IEEE TRANSACTIONS ON MULTIMEDIA and COMPUTER VISION AND IMAGE UNDERSTANDING. He received the Best Paper Awards from ACM MM 2012 (demo), PCM 2011, ACM MM 2010, ICME 2010, and ICIMCS 2009, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC 2010, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, 2012 NUS Young Researcher Award, and the co-author of the best student paper awards of PREMIA 2009, PREMIA 2011, and PREMIA 2012.

Huadong Ma (M'14) received the B.S. degree in mathematics from Henan Normal University, Henan, China, in 1984, the M.S. degree in computer science from Shenyang Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 1990, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 1995.

He visited United Nations University International Institute for Software Technology as research fellow in 1998 and 1999. From 1999 to 2000, he held a visiting position in the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, USA. He was a visiting Professor at The University of Texas at Arlington, Arlington, TX, USA, from July 2004 to September 2004, and a visiting Professor at Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, from December 2006 to February 2007. He is currently a Professor and Director of Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China. His current research focuses on multimedia system and networking, sensor networks, and Internet of Things, and he has published over 100 papers and 4 books in these fields.

Tae-Kyun Kim (M'11) received the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2008.

He was a Junior Research Fellow of Sidney Sussex College, Cambridge, U.K., from 2007 to 2010. He has been a lecturer in computer vision and learning at the Imperial College, London, U.K., since 2010. His research interests span various topics, including object recognition and tracking, face recognition and surveillance, action/gesture recognition, sematic image segmentation and reconstruction, and man-machine interface.

Dr. Kim has co-authored over 40 academic papers in top-tier conferences and journals in the field, 6 MPEG7 standard documents, and 17 international patents. His co-authored algorithm is an international standard of MPEG-7 ISO/IEC for face image retrieval.