Finding Celebrities in Billions of Web Images

Xiao Zhang *, Lei Zhang, Member, IEEE, Xin-Jing Wang, Member, IEEE, and Heung-Yeung Shum, Fellow, IEEE

Abstract—In this paper, we present a face annotation system to automatically collect and label celebrity faces from the web. With the proposed system, we have constructed a large-scale dataset called “Celebrities on the Web” (CFW), which contains 2.45 million distinct images of 421,436 celebrities and is orders of magnitude larger than previous datasets.

Collecting and labeling such a large-scale dataset pose great challenges on current multimedia mining methods. In this work, a two-step face annotation approach is proposed to accomplish this task. In the first step, an image annotation system is proposed to label an input image with a list of celebrities. To utilize the noisy textual data, we construct a large-scale celebrity name vocabulary to identify candidate names from the surrounding texts. Moreover, we expand the scope of analysis to the surrounding texts of webpages hosting near-duplicates of the input image. In the second step, the celebrity names are assigned to the faces by label propagation on a facial similarity graph. To cope with the large variance in the facial appearances, a context likelihood is proposed to constrain the name assignment process. In an evaluation on 21,735 faces, both the image annotation system and name assignment algorithm significantly outperform previous techniques.

Index Terms—Image Retrieval, Image Annotation, Face Recognition, Image Database

EDICS Category: 1-FACE

I. INTRODUCTION

Recent years have witnessed the explosive growth of large-scale image archives on the Web. The availability of this virtually unlimited supply of web images has had a great impact on multimedia research. Among various web images on any topic, celebrity images including portraits, posters, movie snapshots and news images are of particular interest to end-users. The fact that celebrity-related queries constantly rank the highest among all the image queries clearly reveals the intensive user interest for celebrity images. But exactly how such images can be harnessed and organized remains a critical problem. To better serve the end-user demand and foster multimedia research, we propose a scalable and accurate face annotation approach to name celebrities in general web images.

Collecting and labeling celebrity faces from general web images are much more difficult than the tasks in previous works, which mostly focus on labeling celebrity faces in special types of web images, like news [1]–[4] or social network [5]. The major challenges lie in the noises of web data. Firstly, the surrounding text of a web image often comprises of words and phrases lacking of a standard grammar structure. Therefore it is difficult to apply natural language processing techniques to extract celebrity names and estimate the likelihood of a celebrity appearing in the image. Secondly, celebrity faces on the web exhibit large visual variation due to pose, makeup, expression and occlusion caused by sunglasses or fancy hairstyles, as shown in Fig. 1. This layer of “visual noise” imposes great difficulty for associating names with faces by visual analysis.

To address the aforementioned challenges, a two-step approach is proposed. In the first step, a novel image annotation system is proposed to label an input image with a list of celebrities by mining the surrounding texts. To handle the noisy textual information, a large-scale celebrity name vocabulary is automatically constructed from the web to filter the surrounding texts and identify candidate names. Moreover, we expand the analysis to the surrounding texts of the near-duplicates of the input image. The intuition is that, the names of the celebrities who truly appear in the input image tend to occur frequently in the hosting webpages of its near-duplicates. In the second step, we assign the celebrity names in the image annotation results to the faces. To handle the tremendous volume of web images, the assignment is performed in an unsupervised manner. This is achieved by first deriving (weak) face labels based on the image annotation results, and then refine the face labels by propagating information from highly confidence labels to less confident ones on a facial similarity graph. To handle the large variation in the facial appearances, we propose a context likelihood to constrain the label propagation process. In the experiment section, we show that both the image annotation system and the name assignment algorithm achieve significant accuracy improvements over baseline algorithms.

Next, we’ll first introduce the collected large-scale face dataset in Section II, and then discuss related works in Sec-

Fig. 1. Sample faces from automatically collected and labeled CFW database with diversified pose, expression and lighting conditions.
An overview of the proposed web face annotation approach is given in Section IV, followed by discussions on the proposed image annotation system and the name assignment algorithm in Section V and Section VI respectively. Finally, we discuss the evaluation results in Section VII and conclude the paper in Section VIII.

II. CFW DATASET CONSTRUCTION

A large-scale celebrity dataset named “Celebrities on the Web (CFW)” was constructed using the proposed approach. Starting from two billion web images collected from the Internet, 300 million face images were detected using a robust face detector [6], among which 70.8 million images were processed by the proposed face annotation system. To ensure the diversity of the dataset, we discarded near-duplicate images during the data collection process. Following this procedure, a database of 2.45 million celebrity images corresponding to 421,436 celebrities was collected.

The CFW dataset is advantageous in several aspects. First, it is orders of magnitude larger than previous datasets, most of which contain only tens of thousands of faces [7]–[12]. Second, since all faces in CFW are collected from the web, they exhibit large variation in pose, expression, hairstyle and makeup. Examples are shown in Fig. 1. Finally, CFW dataset is much more accurate than previous automatically generated datasets. For example, the error rate of Faces in the News dataset [1] is 23%, while the overall error rate of CFW dataset is 13.93% and a significant portion of CFW dataset (constituting over half of the CFW dataset) achieves an error rate as low as 4.07%. Detailed comparisons between CFW and existing popular face datasets are shown in Section VII-B. Because of its scale, diversity and accuracy, the collected CFW dataset is of great value to both the multimedia research community [4] [13] [14] [1] and many industrial applications [15]–[17].

III. RELATED WORK

Due to the importance of training data for face recognition, many works have been done to create large-scale face databases with name labels. The majority of the existing approaches require human judgment to produce reliable labels, such as [12] and [18], which is difficult to scale up to labeling millions of faces.

In order to create larger scale databases, a number of works attempted to associate names to faces for special types of web images. For example, works in [1]–[4] focused on associating faces in news photos with names. Berg et al [19] [1] applied a named entity detector on the image captions, and a face detector on the images, then tried to find associations between detected names and faces. In their work, faces were represented based on kernel PCA and clustered using a Gaussian mixture model, where each component was assumed to correspond to a specific person. The parameters of the Gaussian mixture model were learnt together with the most probable associations between names and faces. Guillaumin et al [3] developed a graph based methods to resolve name face association. They modeled the name-face association as a bi-partite graph matching problem with names as one set of nodes and faces as another set, and used a min-cost max-flow algorithm to speed up the matching process. Inspired by machine translation methods, Pham et al [4] proposed to model the name face association problem as cross-media alignment by exploring the symmetry between the visual and textual modalities. An interesting idea in their work is to model the probability of a face being described in the corresponding image caption based purely on visual information, e.g., the size and position of the face.

The reason why news photos are favored in above works is because their associated captions are usually well-written so that name labels can be easily extracted. However, this is not true for general web images, the surrounding texts of which are often noisy and lack a rigorous grammar structure. Besides, a common assumption made in above works is that the same person must not appear multiple times in an image. This assumption might be acceptable for news pictures but it is inappropriate for general web images which contains an enormous amount of user edited pictures, as shown in Fig. 2.

Compared with [1]–[4], our approach has several advantages. First, the proposed image annotation system is capable of labeling names to general web images, while name label extraction algorithms in [1] [4] is not suitable for this task as previously discussed. Second, our name assignment algorithm do not impose any assumption on the facial feature distribution, while a gaussian distribution is assumed for the faces of a person in [1], which may be unrealistic for real-world face database. Third, only visual cues are used in [2] [3], which are not sufficient for annotating web celebrity faces not only because of the semantic gap but also a large variation in facial appearances. In contrast, our method combines a context likelihood and visual similarity to improve the assignment accuracy.

Besides generating face dataset from news image, search engines are used in [20] [15] to obtain celebrity images with name labels. However, these name labels are unreliable because the error rate of image search engine increases dramatically for images with lower ranks. And human verification is required to improve the accuracy as in [20]. Further, search engine-based methods can only associate the query names to the returned images, which is insufficient for handling images containing multiple celebrities.
Fig. 3. Celebrity face annotation framework. First, we propose a method to automatically construct a large-scale celebrity name vocabulary from semi-structured web data and then input it to the face annotation system to process a target image, as shown in Fig. 3 (a). Then an image annotation system is proposed to label celebrity names to the target image, as shown in Fig. 3 (b) with example annotation results in Fig. 3 (c). Based on the annotation result (i.e. the image-level names), we obtain weak labels for celebrity faces as in Fig. 3 (d). Finally, a Bayesian framework is adopted to derive soft face labels by combining a context likelihood and visual similarity, as shown in Fig. 3 (e). Note that, with information from multiple webpages, we are able to remove incorrect names like Connor Cruise in the second surrounding text of Fig. 3 (b).

Another related area is image annotation. Recent advances in image annotation have led to large-scale systems capable of annotating very specific information to images, e.g. Kennedy et al [21] and Arista [22]. For example, for a picture of Michael Jackson, while early image annotation prototypes may only label this picture as “human”, Arista is able to generate more specific and informative tags like “Michael Jackson”, “rock star”, “pop king”. Our work is related to Arista [22] in that both works analyze the surrounding texts of the near-duplicates of input images in order to generate accurate image annotation results. The main difference is that the image annotation system in this paper focuses on labeling a special type of tags, i.e. celebrity names. This leads to a substantial difference in the algorithm design. By leveraging a large celebrity name vocabulary to filter surrounding texts and exploring novel features to represent annotation keywords, the proposed algorithm achieves higher name labeling accuracy.

Also related is the large-scale face retrieval work proposed by Zhong et al [23], which enables efficient face retrieval in a repository of millions of faces. Combining the CFW database with such a large-scale face retrieval system will lead to a large-scale real world face recognition engine. We evaluate such a prototype system in the experiment section.

IV. OVERVIEW OF OUR APPROACH

In this section, we provide an overview to the proposed approach for celebrity face annotation, as shown in Fig. 3. The proposed web face annotation system consists of two components. First of all, we propose an image annotation system to determine who might appear in an input image by identifying celebrity names from surrounding texts, as shown in Fig. 3(b) and Fig. 3(c). Then, given a set of names from image annotation results, we propose a name assignment algorithm to assign the names to the faces in the input image, corresponding to Fig. 3(d) and Fig. 3(e). Next, we will give a brief overview to the two components.

A. Image Annotation System

The proposed image annotation system consists of three steps:

1) Construct a large-scale celebrity name vocabulary.
2) Given an input image, discover all webpages hosting its near-duplicates.
3) Use the name vocabulary to filter the surrounding texts of discovered webpages and identify candidate names. Compute confidence scores of candidate names by analyzing their distribution in the surrounding texts.

The motivation is twofold. First, Due to the lack of grammar and structure, textual information on the web is more effectively analyzed by a name vocabulary based approach than methods relying on natural language cues, e.g. named entity recognition. Second, aggregating information from multiple webpages from independent websites helps remove noise and identify names of celebrities who truly appear in the image.

The annotated images can be divided into three groups:

• **SFSN**: images with one face detected and one name label;
• **SFMN**: images with one face detected and multiple name labels;
• **MF**: images with multiple faces detected.

With **SFSN** images, we can derive a large number of accurately labeled faces. This is because the name labels of **SFSN** images usually have high confidence, and since there is only one face in an **SFSN** image, its image-level name can be used directly as its face label. As shown in Section VII-B, the accuracy of face labels for **SFSN** images is 95.93%. However, for **SFMN** and **MF** images, there are no simple ways to derive accurate face labels only from image annotation results. For **SFMN** images, textual information alone is not sufficient to infer their face labels accurately, as shown in the experiment results in Section VII-C. While for **MF** images, we can not determine how to assign names to faces without using visual cues. Since **SFMN** and **MF** images constitute over 42% of
the collected dataset, it is necessary to develop a method to combine textual and visual information to accurately assign their identified names to faces. We will address this problem in the next subsection.

B. Multi-Modal Name Assignment

To assign names to faces for SFMN and MF images, we propagate highly confident labels from SFSN images. Given an input face, the propagation is conducted by performing instance search in the SFSN images and then let the nearest neighbors vote for their labeled names.

Due to the large variation in the visual appearances of celebrity faces, it is very likely that the nearest neighbors include faces from irrelevant individuals. To handle this layer of “visual noise”, a context likelihood is developed to constrain the propagation process. The context likelihood incorporates the information from surrounding texts by using the confidence scores estimated by the image annotation system. Besides, the context likelihood also incorporates an assumption that it is uncommon for the same person to appear multiple times in one image. A Bayesian framework is adopted to combine the visual propagation results and the context likelihood.

Next, we will discuss the image annotation system and name assignment algorithm in detail.

V. IMAGE ANNOTATION SYSTEM

The goal of the proposed image annotation system is to label an input image with a list of celebrities who may appear in the image. An illustration of the annotation system and example results are shown in Fig. 3(b) and (c).

A. Constructing a Large-scale Celebrity Name Vocabulary

As aforementioned, we identify names from surrounding texts with a large-scale vocabulary. We construct such a vocabulary by utilizing the semi-structured data in Wikipedia. The entire Wikipedia contains 5 million pages, out of which 1.7 million pages are considered as articles describing some unique concepts [24]. In this work, we train a classifier to identify people related pages from all Wikipedia pages, and then extract names from their page titles.

To classify a page, we first extract bag-of-word features from the first paragraph of the page, which gives a brief introduction of the subject of the page. Beside the paragraphs in the article, another part of the page, namely information box, is also informative for classification. Information box is a group of property-value pairs which summarize the basic characteristics of the subject in the page. Since different subjects have different properties, this data is highly informative. For example, some common properties in the infobox of a person include “Born”, “Occupation”, “Spouse”, etc.. However, the problem with information box is that it is not available for every Wikipedia page. This drawback is alleviated by incorporating category tags, which are usually provided at the bottom of Wikipedia pages for the purpose of grouping pages with similar topics. Some category tags are very discriminative for identifying people pages, such as “1962 births”, “20th-century actors”, “American computer scientists”, etc. From information box and category tags, we also extract simple bag-of-word features.

In this way, a page is converted to three kinds of feature vectors, including features based on the first paragraph of an article, features based on the information box and features based on the category tags. Support vector machine is used to learn prediction models separately for the three features. After comparing the three features, we found category tags achieved the best performance on our testing data of 3K Wikipedia pages (another 3K for training), with classification accuracy over 95%. Therefore we chose to rely on category tags to identify people pages. After parsing the entire Wikipedia corpus, 750,555 people pages were collected and their page titles were used to construct our name vocabulary.

Besides Wikipedia, we also use Entitycube [25] to construct the vocabulary, which automatically collects entities from general web pages. The accuracy of the name vocabulary from Entitycube is lower but contains a substantial number of less well-known people (8.78 million).

After the celebrity name vocabulary is generated, the universe of possible name labels are defined as

\[ \mathcal{V} = \{ v_k | k = 1, \ldots, K \} \]

where \( K \) is the total number of celebrities in our name vocabulary.

B. Discover Related Webpages by Near-duplicate Image Retrieval

Annotate an input image with information only from its hosting webpage may suffer from the noise in the surrounding text. Therefore we expand the scope of analysis to the surrounding texts of webpages hosting near-duplicates of the input image. The assumption is that, since these webpages come from different websites, aggregating their surrounding texts will magnify the signals from correct names and dilute the noises from incorrect ones.

To perform near-duplicate image retrieval in a web scale database, we collaborated with a product team in Microsoft. Several hundreds of machines were allocated for this task. Although finding near-duplicates for a query image from two billion images seems extremely challenging, the divide and conquer strategy utilizing several hundred machines makes this task possible as each server only needs to deal with a few millions of images. In this way, many near-duplicate image retrieval algorithms [26]–[29] can be employed to perform the task. Since near-duplicate image detection is not the contribution of this paper and due to confidential reasons, we skip the implementation details.

C. Image Annotation by Mining Surrounding Texts of Related Webpages

With the name vocabulary and the expanded set of surrounding texts, a group of candidate names can be identified. Next we discuss how to generate the final annotation results. Denote \( I \) as the image to annotate, \( \{ I_i | i = 1, \ldots, M \} \) as the set of near-duplicate images of \( I \), \( T_i \) as the surrounding texts for \( I_i \).
Denote the image context for \( I \) as \( \mathcal{T} = \{ T_i | i = 1, \ldots, M \} \) and names extracted from \( \mathcal{T} \) as \( \mathcal{V} = \{ v^I_k | k = 1, \ldots, K_I \} \)

Note that \( \mathcal{V} \subseteq \mathcal{V} \), and usually \( K_I \ll K \).

The annotation task is a multi-label problem: to choose a set of names\(^1\) from \( \mathcal{V} \) to annotate \( I \). We solve this problem by learning a binary classifier. Given a name \( v^I_k \), the classifier outputs an image-level name label \( z_k \) as follows:

\[
z_k = \begin{cases} 1 & \text{if a celebrity named } v^I_k \text{ appears in } I \\ 0 & \text{if no celebrity in } I \text{ is named } v^I_k \end{cases}
\] (1)

Denote the features for \( v^I_k \) as \( \mathbf{X}(v^I_k) \). \( \mathbf{X}(\cdot) \) considers the following three aspects:

1. **Type of names**: There are three ways that we consider a term in a surrounding text as a candidate name: 1) full name, e.g., “Harry Potter”; 2) partial name, e.g. “Harry” or “Potter”; 3) concatenated name, e.g. “harrypotter”. In our experiments, we observed that full name match was a very reliable indicator for the occurrences of celebrity names. Partial match can increase the recall of annotation because people often use partial names for simplicity. And concatenated name match is also important for extracting candidate names, especially in image URLs.

2. **Type of surrounding texts**: Surrounding text of an image is the textual information in the hosting webpage related to the content of the hosted image. In our implementation, seven types of surrounding texts are extracted to represent a web image, i.e. caption, alt text, image URL, page title, section title, forty words surrounding the image link, and homepage names. We observed that different kinds of surrounding texts have various importance. For instance, information in image URL or caption is more predictive than information in page title.

3. **Frequency vs. Ratio**: Here, frequency corresponds to the number of times that \( v^I_k \) occurs in \( \mathcal{T} \), while ratio \( r \) measures the percentage of \( \mathcal{T} \) in which \( v^I_k \) occurs. Intuitively, \( v^I_k \) is more likely to correspond to a celebrity appearing \( I \) if both its frequency and ratio are high.

We measure the probability that \( v^I_k \) appears in \( I \) by:

\[
p(z_k = 1 | \mathcal{T}) = \frac{1}{1 + e^{-(\mathbf{W}^T \mathbf{X}(v^I_k) + b)}}
\] (2)

where parameters \( \mathbf{W} \) and \( b \) are learnt by logistic regression. If \( p(z_k = 1 | \mathcal{T}) > 0.5 \), then \( v^I_k \) is labeled as an image-level name for \( I \).

VI. MULTI-MODAL NAME ASSIGNMENT

In this section, we discuss the proposed name assignment algorithm which propagates the highly accurate labels of SFSN images to faces in SFMN and MF images. This algorithm is entirely unsupervised and thus avoids manual labeling in order to process large-scale datasets. In the algorithm, we do not propagate labels between faces in SFMN and MF images to avoid propagating noise.

\(^1\)A name will be included in \( \mathcal{V} \) only if it occurs in the surrounding texts of at least two near duplicate images.

A. Overview of the Assignment Model

Denote the detected faces in image \( I \) as:

\[
\mathbf{F} = \{ f_1, f_2, \ldots, f_P \}
\]

where \( P \) is the total number of faces in \( I \), and each \( f_i \) (\( i = 1, \ldots, P \)) is a feature vector describing the appearance of the \( i \)-th face in \( I \).

Correspondingly, denote the name assignment vector as:

\[
\mathbf{Y} = \{ y_1, y_2, \ldots, y_P \}, y_i \in \mathcal{V} \cup \{ \text{null} \}
\]

where \( y_i \) is the label of \( f_i \) \( (i = 1 \ldots P) \), “null” means no name is assigned to \( f_i \), and \( \mathcal{V} \) is the set of image-level names extracted from surrounding texts as defined in Section V-C.

Denote the confidence for label \( y_i \) of face \( f_i \) as \( \psi(y_i|f_i) \), which is estimated by computing the posterior distribution of \( y_i \) conditioned on the facial features \( \mathbf{F} \), image context \( \mathcal{T} \) as defined in Section V-C and a parameter \( \lambda \):

\[
\psi(y_i|f_i) = p(y_i|\mathbf{F}, \mathcal{T}; \lambda)
\] (3)

where \( \mathcal{T} \) is defined in Section V-C and \( \lambda \) is a parameter of the proposed algorithm. From the definition of Eq. 3, we can see that both visual and textual cues are utilized in the name assignment algorithm.

To compute \( p(y_i|\mathbf{F}, \mathcal{T}; \lambda) \), we marginalize the joint posterior distribution of all the name assignments in image \( I \), i.e. \( p(\mathbf{Y}|\mathbf{F}, \mathcal{T}; \lambda) \). Assuming \( \mathbf{F} \) and \( \mathcal{T} \) are conditionally independent given \( \mathbf{Y} \), the joint posterior distribution becomes:

\[
p(\mathbf{Y}|\mathbf{F}, \mathcal{T}; \lambda) \propto \frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{Y}|\mathcal{T}; \lambda)}{p(\mathbf{Y})}
\] (4)

In Eq. (4), there are three core components: \( p(\mathbf{Y}|\mathbf{F}) \) propagates name labels by visual similarity, \( p(\mathbf{Y}|\mathcal{T}; \lambda) \) is the context likelihood to constrain the propagation process, and \( p(\mathbf{Y}) \) is the prior distribution of name assignments. Next we discuss how to estimate the three components.

B. Label Propagation from SFSN Images \( p(\mathbf{Y}|\mathbf{F}) \)

In this subsection, we discuss how to propagate labels from SFSN images to SFMN and MF images through a facial similarity graph. Denote the set of faces in SFSN images as:

\[
\tilde{\mathcal{F}} = \{ \tilde{f}_n | n = 1, \ldots, N \}
\]

where \( N \) is the total number of SFSN images. By using Eq. (2), the labels for SFSN faces are initialized directly from their image-level names:

\[
\psi(v_k|\tilde{f}_n) = p(z_k = 1 | \mathcal{T}(\tilde{f}_n))
\] (5)

As introduced in Section VI-A, \( \psi(v_k|\tilde{f}_n) \) is the confidence of assigning name \( v_k \) to a SFSN face \( \tilde{f}_n \), and \( z_k \) is a binary variable as defined in Eq. (1) to indicate whether the celebrity named \( v_k \) appears in the image containing \( \tilde{f}_n \). Here, \( \mathcal{T}(\tilde{f}_n) \) is the image context for \( \tilde{f}_n \) as defined in Section V-C.

Names of SFSN faces are propagated to the detected faces \( \mathbf{F} \) in image \( I \) independently. Assume that the name assignment
vector is $Y = (v^I_k, v^I_{k2}, \ldots, v^I_{kP})$, i.e. to assign name $v^I_k$ to $f_i$ where $i = 1 \ldots P$. Then the propagation formula is:

$$p(Y|F) = \prod_{i=1}^{P} p(v^I_{ki}|f_i), v^I_{ki} \in \mathcal{V}^I \cup \{\text{null}\}$$

$$p(v^I_{ki}|f_i) = \sum_{\tilde{f}_i, \in \mathcal{N}(f_i)} \psi(\tilde{f}_i|f_i)p(\tilde{f}_i|f_i)$$  \hspace{1cm} (6)

where $\mathcal{V}^I$ is the set of image-level names for image $I$, $\mathcal{N}(f_i)$ is defined as the set of $k$ nearest neighbors of $f_i$ among SFSN faces, and $p(\tilde{f}_i|f_i)$ is the transition probability from $f_i$ to its neighbor $\tilde{f}_i$, as defined by the following formula:

$$p(\tilde{f}_i|f_i) = \frac{e^{-\frac{d(\tilde{f}_i, f_i)}{L_1}}}{\sum_{\tilde{f}_i, \in \mathcal{N}(f_i)} e^{-\frac{d(\tilde{f}_i, f_i)}{L_1}}}$$  \hspace{1cm} (7)

where $L_1$ distance is used to compute $d(\tilde{f}_i, f_i)$ for efficiency.

The propagation can be viewed as a weighted voting process. To estimate the confidence of assigning name $v^I_{ki}$ to $f_i$, each of its neighbor $\tilde{f}_i$ votes according to the label confidence $\psi(v^I_{ki}|\tilde{f}_i)$, and the votes are weighted by the similarity between $f_i$ and $\tilde{f}_i$.

### C. Constrain the Propagation by a context likelihood $p(Y|T; \lambda)$

In this subsection, we introduce how we model contextual information by $p(Y|T; \lambda)$ in Eq. (4).

Given a name assignment vector $Y$, for each image-level name $v^I_k (k = 1 \ldots K_I)$ in $\mathcal{V}^I$ extracted by Eq. (2), we know whether it appears in the image (indicated by a binary variable $z_k$ as defined in Eq. (1)) and the number of faces it corresponds to (i.e. the number of times $v^I_k$ is assigned, denoted by $m_k$). Here, we want to impose a soft constraint on $Y$ such that $v^I_k$ with a higher confidence estimated by Eq. (2) has a higher probability to be assigned. Also, $m_k$ should have a small probability to be larger than 1 because it is uncommon for the same person to occur multiple times in an image. Therefore, we model the context likelihood as the joint distribution of $z = \{z_k\}$ and $m = \{m_k\}$ given $T$, where $k = 1 \ldots K_I$.

Denote the set of name labels assigned to at least one face in $I$ as $\{v^I_k | l \in 1 \ldots K_I^+\}$, and the rest of the name labels as $\{v^I_k | q \in 1 \ldots K_I^-\}$, where $K_I^+ + K_I^- = K_I$. The context likelihood is modeled as:

$$p(Y|T; \lambda) = p\left(\left\{\left\{z_k, m_k|\lambda\right\}\right\}; \lambda\right)$$

$$= p(z_k|T)p(m_k|z_k; \lambda)$$

$$= \prod_{l=1}^{K_I^+} p(z_{kl} = 1|T) \prod_{q=1}^{K_I^-} p(z_{kq} = 0|T)$$

$$\times p(m_k|z_k; \lambda)$$  \hspace{1cm} (8)

As defined in Eq. (2), $p(z_{kl} = 1|T)$ measures the confidence of the celebrity named $v^I_k$ appearing in $I$ given the image context $T$. $p(z_{kq} = 0|T)$ serves as a punishment for not assigning name $v^I_k$ to any face. $p(m_k|z_k; \lambda)$ is defined as:

$$p(m_k|z_k; \lambda) = \begin{cases} \prod_{q=1}^{K_I^-} \lambda^{m_k_{q} - 1} Z_{\lambda} & \text{when } m_k > 0 \text{ if } z_{kl} = 1 \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (9)

where $Z_{\lambda}$ is a normalization constant and $P$ is the total number of faces in image $I$ as defined in Section VI-A. We set $\lambda$ as $0 < \lambda < 1$ so that $p(m_k|z_k; \lambda)$ is a monotonically decreasing function of $m_k$ where $l = 1 \ldots K_I^+$. In this way, $p(m_k|z_k; \lambda)$ serves as a regularization component to softly punish multiple assignments of the same name, i.e. when $m_k > 1$.

Intuitively, the proposed context likelihood assumes that the faces in image $I$ arise from the following generative process:

1) For each image-level name $v^I_k$, generate a binary variable $z_k$ from $p(z_k|T)$ as defined in Eq. (2) to indicate whether $v^I_k$ appears in image $I$.

2) If $z_k = 1$, generate $m_k$ faces of name $v^I_k$ in image $I$ from $p(m_k|z_k; \lambda)$ as defined in Eq. (9).

Note that this is different from the hard constraints imposed by previous works [1] [3] [30] requiring that the same name must not be assigned to multiple faces in an image, which is not valid for general web images because there exist a significant number of manually edited web images with multiple faces of the same celebrities, as previously shown in Fig. 2.

---

**Fig. 4.** An example is provided to illustrate the effectiveness of the context likelihood. On the left of the figure is the image annotation results. In the middle of the figure, we show how context likelihood helps correct name assignment error caused by noises in the visual similarity measurement. Here, the noises refer to the large variation in visual appearances of celebrity faces and lead to incorrect nearest neighbors as shown on the right of the figure.
Fig. 4 illustrates the power of the context likelihood. On the left of the figure, we show that the image is annotated with two names, which are “Angelina” with confidence 0.65 and “Clint” with confidence 0.60. On the right, it is shown that the algorithm relying solely on visual similarity incorrectly assigns “Angelina” to both faces. This is because there are much more SFSN faces of “Angelina” than “Clint”, creating a class imbalance problem for the label propagation. In the middle, we can see that incorporating the context likelihood leads to a significant accuracy boost.

D. Normalization by Name Prior $p(\mathbf{Y})$

In Eq. (4), $p(\mathbf{Y})$ represents the prior of names. This prior is estimated using the labels of SFSN images. Assume that the name assignment vector is $\mathbf{Y} = (v^1_f, v^2_f, \ldots, v^{P_f} f)$ and that all of the individual dimensions of $\mathbf{Y}$ are independent. We have:

$$p(\mathbf{Y}) = \prod_{i=1}^{P} p(v^i_f)$$

(10)

Assume that the distribution of facial descriptors of SFSN faces is uniform, and that the prior distribution of name $v^i_f$ can be estimated from the label confidences $\psi(v^i_f | f_j)$ of SFSN faces $f_j$ ($j = 1 \ldots N$) as defined in Section VI-B, the formula to compute $p(\mathbf{Y})$ is then given by:

$$p(v^i_f) = \sum_{j=1}^{N} \psi(v^i_f | f_j)/N$$

$$p(\mathbf{Y}) = \prod_{i=1}^{P} \sum_{j=1}^{N} \psi(v^i_f | f_j)/N$$

(11)

where $P$ is the total number of faces and $v^i_f$ is the name assigned to the i-th face in image $I$, $N$ is the total number of SFSN faces in our dataset.

Adding a name prior is meaningful not only from a Bayesian perspective but also for performance issue. Without such a normalization, popular names will have better chances to be propagated from SFSN faces to SFMN or MF faces, which is not desired.

E. Implementation Detail: Face Representation

In this work, the appearance of each face is described by Local Binary Pattern (LBP) [31], which is widely used in face recognition [32][36]. In our implementation, the face image is divided into small regions from which the LBP features are extracted and concatenated into a single feature histogram. The idea behind the LBP features is that the face images can be seen as composition of micro-patterns which are invariant with respect to monotonic grey scale transformations. This is of important value when handling faces under a large variety of illumination conditions. Although alternative facial descriptors [19] [37] [38] exist, a thorough comparison is beyond the scope of this paper.

To make facial similarity computation efficient on millions of faces, we apply PCA [39] to reduce the dimension of face descriptor from over 3,000 dimensions to 500 dimensions. The number of dimensions is chosen to balance the computation efficiency and the representation power.

VII. Evaluations

In this section, we evaluate the proposed algorithm, the constructed database, and its application to celebrity recognition.

A. Groundtruth Dataset Construction

To evaluate the performance of the proposed algorithm and the accuracy of the automatically annotated CFW database, we manually labeled 21,735 faces randomly sampled from the CFW database. Among these faces, 500 of them were used to train the logistic regression model in Eq. (2). The rest made up of the groundtruth dataset for all the following evaluations. The detailed information about the groundtruth dataset is summarized in Table I.

To label the groundtruth dataset, we partitioned the faces into 7 subsets and invited 7 volunteers for the labeling work. Each subset of faces were labeled by two volunteers. If the two volunteers provided different labels, extra human judge was assigned to determine the appropriate labels. To facilitate the labeling process, candidate names were proposed from our image annotation algorithm. If none of the candidate names were the correct names, “null” were labeled to the corresponding faces. When volunteers encountered an unknown celebrity, they were required to look at top Bing and Google image search results to get familiar with the visual appearance of the celebrity. Moreover, the faces were grouped by their candidate names, which greatly sped up the labeling process. On average, each volunteer spent 10 hours on the labeling work.

B. Properties of CFW Database

In this section, we evaluate the scale and accuracy of the CFW dataset.
**TABLE III**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Images</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFW-SFSN</td>
<td>1,429,878</td>
<td>4.07%</td>
</tr>
<tr>
<td>CFW</td>
<td>2,453,402</td>
<td>13.93%</td>
</tr>
<tr>
<td>Faces in the Wild [19]</td>
<td>30,281</td>
<td>23.00%</td>
</tr>
<tr>
<td>Labeled Yahoo News [30]</td>
<td>28,204</td>
<td>20.58%</td>
</tr>
</tbody>
</table>

**TABLE IV**

<table>
<thead>
<tr>
<th>#SFSN Faces</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,429,878</td>
<td>4.07%</td>
</tr>
<tr>
<td>857,927</td>
<td>2.71%</td>
</tr>
<tr>
<td>571,951</td>
<td>2.23%</td>
</tr>
<tr>
<td>428,963</td>
<td>1.95%</td>
</tr>
<tr>
<td>214,481</td>
<td>1.69%</td>
</tr>
</tbody>
</table>

**Scale:** The CFW dataset contains 2.45 million distinct images (i.e., near-duplicate faces are removed in the database) of 421,436 individuals. Here we compare the scale of CFW database with popular large-scale face datasets in Table II. From the table, it can be observed that the CFW database is much larger than previous face databases in terms of both individuals and images.

**Accuracy:** In Table III, we compare the accuracy of CFW with two other automatically labeled face datasets - Faces in the News [19] and Labeled Yahoo News [30]. We can see that the scale and accuracy of our dataset significantly outperform both Faces in the News and Labeled Yahoo News. Besides, we can see that a significant portion of the CFW database is comprised of SFSN images with 95.93% label accuracy, which is far ahead of the results in previous works. Moreover, we can apply a variety of thresholds on the annotation confidences of SFSN faces to get error rates as low as 1.69% while still including a large number of faces, as shown in Table IV.

An interesting observation is that the number of images for a celebrity follows a long tail distribution, as shown in Table V. From the table, it can be seen that the image distribution in the database is highly imbalanced. 64,542 celebrities have more than five images, while only 434 of them have over 500 images. The imbalance nature of the database makes it more realistic and more challenging to existing face recognition algorithms. Finally, we list the top five celebrities with the largest numbers of distinct images in CFW database in Table VI.

**C. Image Annotation**

In this section, we evaluate the proposed image annotation algorithm for generating image-level names (IA-ILN). To evaluate the annotation results, we check whether the $P$ most confident image-level names appear in an image of $P$ faces.

In the first experiment, we compare our image annotation algorithm with three baseline algorithms based on the distribution of celebrity names in a single webpage [40]. The result is summarized in Table VII. It can be seen that, by leveraging information from multiple webpages, our algorithm achieves much better results than the baseline algorithms.

To further demonstrate the effectiveness of using multiple webpages discovered by near-duplicate image detection, we conduct a second experiment by varying the number of webpages discovered by near-duplicate image detection, the proposed algorithm can effectively suppress the noise in the surrounding texts. Another interesting observation is that the annotation error rate of SFSN images drops much faster than those of SFSN and MF images when increasing the number of webpages used in the algorithm.

**D. Multi-modal Name Assignment**

In this section, the proposed multi-modal name assignment algorithm (MMNA) is evaluated.

Each of SFSN images has a single image-level name which can be used as face label directly. While for SFSN and MF images, MMNA is used to assign their image-level names to the faces in the images. To evaluate MMNA, all SFSN...
and MF images from the groundtruth dataset are used, which contain 8,132 faces in total.

We compare MMNA with two widely used methods for name assignment. The first one proposes an efficient algorithm based on min-cost maxflow graphs (MaxFlow) [3] to identify the optimal name-face assignment, which is the state-of-the-art algorithm on associating names with faces on the web. To implement the algorithm in [3], two graph construction methods are used. One is KNN linear, where a node in the graph only connects with its K nearest neighbors. The edges between a node and its nearest neighbors are assigned a series of weights: \( k, k-1 \) and so on down to 1. The other graph construction method is \( \epsilon \)-neighborhood method, where non-zero edge weights are assigned only to neighbors whose distances are smaller than \( \epsilon \). The second baseline algorithm is a generative model proposed in [1] which leverages both visual and textual cues, and assumes a gaussian mixture model for facial features with each component corresponding to all the faces of a celebrity.

**Results:** The results are summarized in Table VIII. It can be seen that without the context likelihood (as defined in Section VI-C), the proposed name assignment algorithm achieves moderately better results than baseline algorithms. However, after incorporating context likelihood, the proposed multimodal name assignment significantly outperforms baseline methods, which demonstrates the power of context likelihood.

**Discussion:** The reasons for the relatively high error rate are twofold. First the number of labeled faces for less popular celebrities is limited, as demonstrated in Table IX. From the table, it can be seen that the proposed algorithm encounters much more failures when ground truth names have fewer SFSN faces. Second, we are handling uncontrolled real-world faces. Fig. 6 shows some test faces which are non-frontal faces or occluded. These faces are much more difficult to recognize than facial images taken in a controlled environment, providing novel challenges and opportunities to face related research.

### E. Celebrity Recognition with CFW Dataset

In this section, we evaluate the power of CFW database on celebrity recognition.

**Recognition Approach:** We adopt an instance-search approach for the celebrity recognition task. For a given query face \( f_q \), its nearest neighbors \( \{f_j\} \) in the entire CFW database vote for their labeled names with the following formula:

\[
Score(v_k) = \sum_j \psi(v_k|f_j)p(f_j|f_q)
\]  

(12)

where \( v_k \) is a celebrity name, \( \psi(v_k|f_j) \) is the confidence of face \( f_j \) being named \( v_k \) and \( p(f_j|f_q) \) is the transition probability from \( f_q \) to \( f_j \) as defined in Eq. (7). The name with the highest voting score is returned to the user. Note that the input face is not associated with any image context so no context likelihood can be leveraged.

**Query Images:** A query dataset of 4,737 facial images for 240 celebrities was downloaded from Internet. The 240 celebrities were randomly sampled from the individuals with at least 50 images in the database. The images were downloaded from top Google image search results. Then faces were extracted from the downloaded images by the same face detector [6] for generating CFW dataset. Finally, we manually checked the query faces and their labels to ensure the correctness.

**Evaluation Metric:** We use Recall@K as our evaluation metric.
metric, which measures whether the correct celebrity name can be returned in the top K names ranked by Eq. (12). This measurement is equivalent to recognition rate when K is set to 1.

**Results and Observations:** The evaluation results are summarized in Fig. 7. On average, our system achieves an recognition rate of 85.83%. When considering returned names up to the third highest vote, the probability that a correct name is returned increases to 93.05%. This result is fairly promising, considering that our database contains 421K individuals and 2.45 million images, and there are large variations in pose, expression and illumination of the testing images.

To further challenge the recognition system, in the returned similar faces, we ignore top ones whose visual distances to the testing faces (i.e. d(\(f_q, f_k\)) in Eq. 7) are smaller than a threshold (TH). This most probably will reduce the number of near-duplicates of testing faces in our database. The evaluation results when TH is set to 2000 are also shown in Fig. 7. We can see that even in this challenging situation, a recognition rate of 83.66% is still achieved, which further demonstrates the effectiveness of the developed recognition system as well as the scale and diversity of the CFW dataset.

Finally, we provide a few examples of testing faces and the recognition results in Fig. 8. From these examples, one can observe that pose and expression have a large impact on visual similarity. This indicates that, in order to build an accurate instance-search based celebrity recognition engine, it is crucial to collect a database that includes faces of different visual variations for all the celebrities.

**VIII. CONCLUSION**

In this paper, we have presented an approach to construct a large-scale labeled celebrity face database from web images. The approach comprises of two components. The first component is an image annotation system. By analyzing surrounding texts of the near-duplicates of an input image, the proposed image annotation system is able to provide a set of names corresponding to the celebrities appearing in the image. With the annotation results, a multi-modal name assignment algorithm is proposed to assign names to faces in the image, which is the second component of our approach. The assignment algorithm adopts a probabilistic approach which combines information from both visual features and image context.

Using the proposed approach, a celebrity database of 2.45 million distinct images and 421,436 celebrities is constructed. A celebrity recognition system has been built and evaluated to demonstrate the potential of the database.

In the future work, we will further improve the accuracy of face labels by investigating the visual consistency within the group of faces for a celebrity. Besides, we will try to merge the faces of the same person but with different names, i.e. stage name, nickname, birth name, etc. Meanwhile, we will continue enlarging the celebrity database. Once enough web images are processed, and the accuracy of the labels is further improved, the CFW dataset will be delivered to research communities by making it publicly available online. We hope CFW will become a central resource for a broad range of face related research, e.g. face recognition, clustering, gender recognition, naming celebrities in videos, etc.

**REFERENCES**


### Table 1: Example Results for Celebrity Recognition Prototype

<table>
<thead>
<tr>
<th>Query Face Image &amp; Groundtruth Name</th>
<th>Similar Faces in CFW Database with Annotation Generated By Proposed Algorithm</th>
<th>Recognized Names With Confidence Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emma Watson</td>
<td>Heidi Montag Emma Watson Anne Hathaway Emma Watson Emma Watson Emma Watson</td>
<td><strong>Emma Watson</strong> 25.9%</td>
</tr>
<tr>
<td>Ringo Starr</td>
<td>Ringo Starr Ringo Starr Ringo Starr Brede Hangeland Damien Rice Nathaniel Bartlett</td>
<td><strong>Ringo Starr</strong> 60.9%</td>
</tr>
<tr>
<td>Bill Gates</td>
<td>Youssef Chahine Bill Gates Bill Gates Bill Gates Bill Gates Bill Gates</td>
<td><strong>Bill Gates</strong> 43.2%</td>
</tr>
<tr>
<td>Maria Sharapova</td>
<td>Debra Messing Maria Sharapova Alan Tam Maria Sharapova Maria Sharapova Maria Sharapova</td>
<td><strong>Maria Sharapova</strong> 35.4%</td>
</tr>
<tr>
<td>Will Smith</td>
<td>Tom Cruise Will Smith Will Smith Will Smith Will Smith Will Smith</td>
<td><strong>Will Smith</strong> 67.6%</td>
</tr>
<tr>
<td>Angela Merkel</td>
<td>Pete Rollins Joan Morris Adriana Lima Britney Spears Angela Merkel</td>
<td><strong>Angela Merkel</strong> 21.7%</td>
</tr>
<tr>
<td>David Beckham</td>
<td>Juan Jose Lawrence Dallaglio Tom Craddock</td>
<td><strong>David Beckham</strong> 61.7%</td>
</tr>
</tbody>
</table>

Fig. 8. Example results for celebrity recognition prototype. The prototype adopts an instance-search based approach to recognize input faces. In the figure, the testing faces are shown on the left. Their nearest neighbors retrieved from CFW database are shown in the middle. On the right are the recognized names with confidence scores computed by Eq. 12.

---