Improving Automatic Name-Face Association using Celebrity Images on the Web

Zhineng Chen[1], Bailan Feng[1], Chong-Wah Ngo[2], Caiyan Jia[3], Xiangsheng Huang[1]

{zhineng.chen, bailan.feng}@ia.ac.cn, cscwngo@cityu.edu.hk, cyjia@bjtu.edu.cn, xiangsheng.huang@ia.ac.cn

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Department of Computer Science, City University of Hong Kong, Hong Kong, China
[3] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

ABSTRACT

This paper investigates the task of automatically associating faces appearing in images (or videos) with their names. Our novelty lies in the use of celebrity Web images to facilitate the task. Specifically, we first propose a method named *Image Matching* (IM), which uses the faces in images returned from name queries over an image search engine as the gallery set of the names, and a probe face is classified as one of the names, or none of them, according to their matching scores and compatibility characterized by a proposed *Assigning-Thresholding* (AT) pipeline. Noting IM could provide guidance for association for the well-established *Graph-based Association* (GA), we further propose two methods that jointly utilize the two kinds of complementary cues. They are: the early fusion of IM and GA (EF-IMGA) that takes the IM score as an additional information source to help the association in GA, and the late fusion of IM and GA (LF-IMGA) that combines the scores from both IM and GA obtained individually to make the association. Evaluations on datasets of captioned news images and Web videos both show the proposed methods, especially the two fused ones, provide significant improvements over GA.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

name-face association, celebrity image, multimedia fusion.

1. INTRODUCTION

With the overwhelming volume of people-related images and videos available on the Web, the searching and browsing of multimedia archives by person names is becoming increasingly popular to end users. A key issue to enable such services is how to find the images (or videos) regarding to each person. A common technique toward this target is text-based name matching, which associates an image with a particular person if his/her name appears in surrounding context of the image like captions, subtitles, tags, etc. This technique is efficient as there are huge quantities of multimedia documents appearing together with names. However, this is hard to say as an effective solution, because the appearance of people and their names are not always consistent, or saying, an image with a particular name does not imply the person would appear in it. As a result, the found images could very likely contain many irrelevant results.

Visual-based refinement on the faces extracted from images returned from the name matching is shown to be a useful way to remove irrelevant results. The task is also referred to as name-face association [1-3] in the literature. Over decades many techniques have been proposed for this task in the domains of new videos [4], TV series [5], movies [6-7], news images [2, 3, 8] and Web videos [9-11]. Basically, these techniques differ in the way of how the refinement is formulated, and most of them heavily depend on characteristics of the domains and contextual cues available. Nevertheless, among them the *Graph-based Association* (GA) proposed by Guillaumin *et al*. [2] gives rise to our extra attention. In this work, with the assumption that the faces of a person should exhibit higher visual similarity, the problem of name-face association is converted to identifying densely connected sub-graphs corresponding to the names. The work is appealing for its unsupervised nature, and it is capable of scaling up to large dataset containing many people and faces. However, GA without essential knowledge about people such as how a particular person looks like is somehow difficult and is probably to generate noisy results.

On the other hand, there are studies leveraging online images to disambiguate people. In [9], celebrity face models are learnt by automatically crawling online photos. The Cast2Face approach is proposed in [6], where the movie cast is used to retrieve a set of faces for each actor using Google image search, and then face tracks in the movie are identified as the actor with the lowest reconstruction error under a multi-task joint sparse representation framework. The work by [10] also shows that celebrity faces from Google are useful for face tagging. Celebrity images available online could be seen as vivid cues in understanding people.

In this paper, we study the issue of automatic name-face association by exploiting celebrity images on the Web. To this end, we first propose a method named *Image Matching* (IM) to implement the matching between face and name, where the name is restricted to a few ones that the face is possibly assigned to and is represented by a set of images indexed by Google. To ensure the matching can compatible with common constraints made for name-face association [2-3], a pipeline named *Assigning-Thresholding* (AT) is also proposed. Observing the similarity captured by IM can be viewed as the probability of a probe face belonging to a particular

ICMR'15, June 23 – 26, 2015, Shanghai, China.
Copyright 2015 ACM 978-1-4503-3274-3/15/06 …$15.00
http://dx.doi.org/10.1145/2671188.2749401

person. It is different with that captured by GA. We further propose to fuse the two kinds of complementary cues to benefit the association. Specifically, we consider both the early and late fusion of IM and GA. The early fusion version, termed as EF-IMGA, takes the IM score as an additional information source, which provides weak supervision could be interpreted as who looks more like the probe face, to help the association in GA. In contrast, the late fusion version, termed as LF-IMGA, directly combines the scores from both IM and GA obtained individually to make the association. We evaluate the proposed three methods on both captioned news images and Web videos. Experimental results show that all the methods improve the performance of name-face association. Moreover, the two fused schemes can surpass the GA results by a large margin under different conditions.

2. THE APPROACH

2.1 Preliminaries

Name-face association can be seen as the following challenge: given an image (or video) dataset with faces and names, establish the latent associations between the faces and names in each image. It is generally seen as a constrained optimization problem, which should satisfies the following constraints: (a) a face can associate with at most one name appearing in its surrounding context, or null if no corresponding name is found (null assignment), and (b) a name can be assigned to at most one face in an image.

Since our work is closely related to the well-established *Graph-based Association* (GA) [2], we first give a short review of the method. Given a set of images with faces and names, GA constructs a similarity graph with all the faces as nodes and visual similarities between faces as weights of edges. The objective is to find all sub-graphs corresponding to the names. In the implementation, each sub-graph is initialized with all faces in images containing a particular name. By doing this the sub-graphs may overlapped and both constraints above are temporarily relaxed. Then, GA optimizes name-face assignments image-by-image, where the similarity between probe face f_i^p and sub-graph S_j corresponding to name n_j, denoted as SGA_{ij}, is computed by

$$SGA_{ij} = \frac{1}{T}\sum_{t=1}^{T} exp(-d^2(\boldsymbol{v}_i^p, \boldsymbol{v}_t^j)/\sigma) \qquad (1)$$

where $\boldsymbol{v}_i^p$ and $\boldsymbol{v}_t^j$ are face descriptors of the probe face and the t-th face, i.e., f_t^j, in all T faces of sub-graph S_j, respectively. σ is the heat kernel parameter, $d(\cdot,\cdot)$ is the Euclidean distance.

By using Eq.1, each face is assigned only to the most similar sub-graph. After all images have been optimized, the constraints are correctly enforced. Since sub-graphs are altered in the optimization, recalculation of Eq.1 would generate different similarities thus affects the assignments previously made. This optimization process is repeated iteratively until a fixed number of times are reached or it is converged, i.e., all sub-graphs are not altered in two consecutive iterations. Based on the optimized results, name-face associations are revealed accordingly. Readers are referred to [2] for more details.

2.2 Image Matching

With the observation that face photos of celebrities can be easily searched from the Web, it becomes a feasible way of using celebrity Web faces to distinguish who is destination person from a few candidates without human intervention. Generally, for a probe face f_i^p, let $\mathcal{N} = \{n_1, n_2, \cdots, n_M\}$ be the M candidate names and $\mathcal{F}_j = \{f_1^j, f_2^j, \cdots, f_K^j\}$ be the K faces in the top ranked images returned from Google image search by using n_j as the query. The similarity between probe face f_i^p and name n_j is computed by

$$SIM_{ij} = \min_{1\le t\le K}\left\| exp(-d^2(\boldsymbol{v}_i^p, \boldsymbol{v}_t^j)/\sigma)\right\| \qquad (2)$$

where $\boldsymbol{v}_t^j$ is the face descriptor of the t-th face in $\mathcal{F}_j$, K controls the number of Web faces taken into account and whose influence will be discussed in the experiment section.

Using Eq.2, we obtain the similarity between every probe face and every name. Thus, a naive way to implement the association seems to be assigning name with the highest SIM_{ij} to each face. However, this strategy would violate the common constraints of name-face association, i.e., unable to make the null assignment and assigning a name to more than one face in an image.

To settle the two problems, we propose a 2-step pipeline named *Assigning-Thresholding* (AT). AT first seeks the optimal name-face assignments at image (or video frame) level, and then thresholds on the obtained similarities to decide null assignments. Specifically, in image-level assignment, we firstly collect the n faces and m names corresponding to an image. Secondly, similar to the idea introduced by [2], a bipartite graph with $n+m$ nodes on both sides is constructed to enable matching types of name-face, name-null and face-null. Thirdly, the problem of bipartite graph matching is efficiently solved by applying the Hungarian algorithm on a corresponding cost matrix, from which optimal assignments of the image are derived.

How to set the cost matrix is an essential step to the success of the assignment. In [2], for costs of real name-face pairs, they are set to negative similarity of the name and face, e.g., $-SGA_{ij}$. For costs of those name-null and face-null pairs, the former ones are set to 0 while the latter ones are set to a constant threshold value that serves as the sentinel to null assignments. That is, smaller costs of face-null pairs will generate more number of null assignments. However, as claimed by [3], setting threshold via this way is not intuitive thus increase the difficulty of determining null assignments. Therefore, we keep the costs for real name-face pairs intact and set the costs for name-null and face-null pairs, respectively defined as $cost\_mf$ and $cost\_nf$, to:

$$cost\_mf = -\frac{m\cdot s_{aver}}{2(n+m)} \text{ and } cost\_nf = -\frac{n\cdot s_{aver}}{2(n+m)} \qquad (3)$$

where $s_{aver} = \sum_{i=1}^{n}\sum_{j=1}^{m} SIM_{ij}$ is the average IM similarity between the n faces and m names. This setting benefits the assignment from two aspects. First, it takes both the similarity and the distribution of name and face within an image into account. For example, given an image with large m and small n, Eq.3 is more likely to make name null assigned and less likely to make face null assigned, which is in accordance with common sense. Second, by allocating relatively large (negative) costs to those name-null and face-null pairs, the Hungarian algorithm is less likely to assign face and name to null unless the evident is confident enough. Thus in most cases, a face would associate with a name compatible with the constraints, and their similarity, saying SIM_{ij}, can be viewed as confidence of the assignment.

The *Assigning* step does not address the problem of null assignments. We thus further employ a *Thresholding* step to determine whether an assigned name-face pair should be confirmed as a true association or not. Null assignment is declared if the similarity of a name-face pair, i.e., SIM_{ij}, is below an empirically set threshold θ. Preliminary results show that the AT pipeline could lead to 3%-7% improvements over the raw implementation [2].

In IM, an issue worth mentioning is the correctness of the faces in top ranked images from Google. Generally, the correctness is high for well-known celebrities like those in this paper. But the assumption not always holds for less famous celebrities. Thus, we plan to apply some filtering techniques to remove the noisy faces before feeding them into IM. This will be discussed elsewhere.

2.3 EF-IMGA

The similarity captured by IM can be seen as a kind of external information source that tells us how similar a probe face looks like a particular person. In contrast, the similarity captured by GA, e.g., SGA_{ij}, is computed based upon all or a part of faces in images having the name. Compared with SIM_{ij}, SGA_{ij} is more noisy and less supervised especially those obtained from the first few iterations. Motivated by this, we propose to fuse the two kinds of similarities to benefit the association.

We first discuss the early fusion of IM and GA, i.e., EF-IMGA. EF-IMGA takes the IM similarity as an additional source in image-level optimization of GA, aiming at facilitating the name-face assignment. It fuses the two kinds of similarities using:

$$SEF_{ij} = \alpha \cdot SIM_{ij} + (1 - \alpha) \cdot SGA_{ij} \qquad (4)$$

where α is the weight to generate the fused EF score, i.e., SEF_{ij}.

Fusing the IM and GA similarities using Eq.4 would benefit the association as follows. For GA, the incorporated IM similarity would guide the EF score more targeted toward the name should be assigned, by which some assignments made by inaccurate GA scores could be rectified. For IM, taking into account the GA similarity can reduce the bias of external data, as the IM similarity is obtained independent of the images to be associated.

With the EF score obtained by Eq.4, the proposed AT pipeline is applied to produce the association results. Note that the difference of IM and EF-IMGA lies in that EF-IMGA uses SEF_{ij} in Eq.4 rather than SIM_{ij} in Eq.2 to perform the association.

2.4 LF-IMGA

We then discuss the late fusion of IM and GA, i.e., LF-IMGA. In LF-IMGA, IM and GA provide their decisions, namely s_{ij}^{IM} and s_{ij}^{GA}, on every probe face and every probable name individually in advance. The fused LF score, i.e., SLF_{ij}, is then obtained by:

$$SLF_{ij} = \beta \cdot s_{ij}^{IM} + (1 - \beta) \cdot s_{ij}^{GA} \qquad (5)$$

where β is a weight linearly fused the two kinds of scores.

In Eq.5, we set $s_{ij}^{IM} = SIM_{ij}$ as the external Web images remain intact. To set s_{ij}^{GA}, we first use SGA_{ij} derived from Eq.1 plus with our AT pipeline to perform the association. It is an iterative process that ends with partitioning the whole face graph to many subgraphs. Based on this, s_{ij}^{GA} is obtained by calculating the similarity between probe face f_i^p and sub-graphs S_j. Note that what we interested here is the score rather than the association decision.

Table 1. FP and FA of IM for different K at an FR of approximately 0.5 on LYN and WC.

		K=10	*K*=20	*K*=30	*K*=all
LYN	FA	0.5937	0.6062	0.6106	0.6124
	FP	0.7579	0.8045	0.8202	0.8245
WC	FA	0.5831	0.6077	0.6095	0.6097
	FP	0.5738	0.6584	0.6659	0.6684

Note: *K*=all means all faces from the top 64 images are considered.

With the LF score obtained by Eq.5, the proposed AT pipeline is also used to produce the association results. EF-IMGA and LF-IMGA differ in the way of how the GA similarity is calculated and when the two kinds of similarities are combined.

3. EXPERIMENTS

3.1 Dataset and experimental setup

We use two datasets constructed in real life conditions to evaluate the proposed methods. The first one is the *test* set of *Labeled Yahoo! news* (LYN), which has 9,362 captioned news images of the 23 most frequently occurred people and their friends [2]. The dataset contains a total of 14,827 faces and 1,071 different people. The other one is the *core* set of *WebV-Cele* (WC), which contains 3,194 videos with 41,228 faces and 144 celebrity names [11]. The two datasets are of different media types and therefore pose different challenges. Faces in LYN, despite captured "in the wild", are official news pictures that are often sharply focused and in high resolution, which stand for a relative high quality testbed. On the other hand, faces in WC are extracted from Web videos. They are less likely to be captured well and are often with low resolution. They thus stand for a more challenging testbed.

We evaluate the methods using three metrics measured the performance of name-face association at the face level, namely Face Accuracy (FA), Face Precision (FP) and Face Recall (FR). FA is the fraction of correctly associated faces (including null assignments) over all the detected faces. FP is the same as FA, except that null assigned faces are not included for evaluation. FR calculates the fraction of correctly associated faces over all the labeled celebrity faces. In all the experiments, we omit faces that are labeled as "_notaface" (in LYN) or "hard-to-determine" (in WC). As the face descriptor, we use the 1937-dimensional feature vector extracted from 13 facial regions [5].

3.2 Evaluations

3.2.1 Influence of K in IM

We first evaluate the influence of K in IM. The parameter K determines how many faces in top ranked images from Google are evaluated. We set K to different numbers (maximal 64). The results are listed in Tab.1. To make the results comparable, we adjust thresholds such that the methods all have an FR of around 0.5.

As can be seen, the performance steadily improves with the increase of K, showing that more Web facial images is helpful in characterizing probe faces from both news images and Web videos. We also observe that the results for LYN are significantly better than those for WC. That is in accordance with expected, as matching between image and video face is more challenge than matching within image faces. When K reaches to 30, the improvements are marginal. Since the computational cost of IM grows linearly with K, we fix $K = 30$ in the following evaluations.

3.2.2 Sensibility analysis of parameterα and β

We then analyze the sensibility of parameter α in Eq.4 and β in Eq.5, which are weights in early fusion (EF) and late fusion (LF), respectively. Specifically, we range the parameters from 0 to 1.0 with an increasing step of 0.1. The results are given in Fig.1, where EF-IMGA (LF-IMGA) on both datasets is presented in the left (right) figure. Thresholds are also adjusted such that all results are w.r.t. an FR of around 0.5. It is seen that setting both parameters to 1, i.e., the IM case, perform better than setting them to 0, i.e., the GA case. The optimal performance is achieved by setting both parameters to 0.7 for LYN and 0.6 for WC. In this setting,

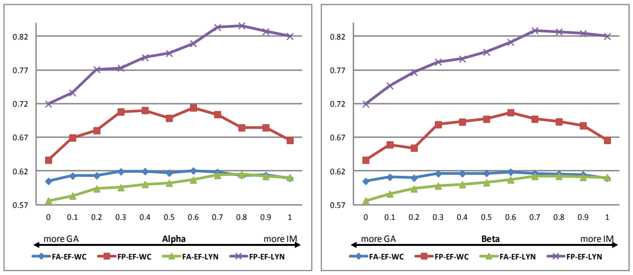

Figure 1: Performance of EF-IMGA and LF-IMGA on both LYN and WC w.r.t. parameters α (left) and β (right).

both EF-IMGA and LF-IMGA perform better than IM on both datasets especially for WC, implying that the name-face association indeed benefits from the fusion of the two kinds of information. The relatively large improvements for EF-IMGA (LF-IMGA) compared with IM on WC are likely to be interpreted as: one on hand, although IM between image and video face are not as straightforward as within image faces, it provides cues that are more different and thus strongly complementary to the cues of GA. On the other hand, the performance of IM on LYN is around 0.82, leaving only a relatively small room to be improved.

3.2.3 Name-face association results

Based on above observations, we compared the proposed IM, EF-IMGA and LF-IMGA with GA [2], to quantitatively analyze the improvements gained by using celebrity images on the Web.

Fig.2 shows FA-FR and FP-FR curves of the four methods on both datasets. The curves are obtained by calculating FA, FP and FR based on different thresholds. The performance of text-based name matching, i.e., FP-Text, is also plotted for reference using horizontal dotted lines. As anticipated, FA-FR curves consistently drop with the growth of FR. However, FP-FR curves experience a first growth and then decline process. This can be explained as the effect of null assignments. When FR is low, many celebrity faces are null assigned because of rigorous set thresholds. The errors are gradually rectified with the growth of FR. When FR becomes high, it causes a new problem of a majority of unknown faces are also assigned with names. Another interesting observation is IM along also performs better than GA even on WC, implying face appearance "in the wild" is so diverse that even IM between heterogeneous domains sounds to be a more stable choice.

We also summarize the results with FRs of around 0.2, 0.5 and 0.8 on both datasets in Tab.2. They are regarded as three typical scenarios of high, middle and low FAs corresponding to applications with different goals. Comparing the two fused methods, EF gives slightly better results in most cases, implying fusing at the feature level, i.e., combining the two kinds of similarity to iteratively guide the name-face assignment in AT, is more reasonable than at the decision level. It is also observed that the two fused methods perform better than GA by a large margin. For example, by using EF and LF, the maximum improvements on both datasets are up to 10.2% and 9.6% in terms of FA, and 16.1% and 15.2% in terms of FP, respectively. The improvements clearly validate the effectiveness of exploiting celebrity images on the Web.

Table 2. Performance of the four methods w.r.t. different levels of FRs on LYN and WC.

		FA			FP		
		FR=0.2	FR=0.5	FR=0.8	FR=0.2	FR=0.5	FR=0.8
LYN	**GA**	0.5133	0.5762	0.5587	0.8868	0.7197	0.5836
	IM	0.5168	0.6106	0.6091	0.9121	0.8202	0.6457
	EF	**0.5189**	**0.6151**	**0.6156**	**0.9404**	**0.8359**	**0.6568**
	LF	0.5178	0.6124	0.6123	0.9336	0.8288	0.6525
WC	**GA**	0.5782	0.6051	0.5446	0.8279	0.6365	0.4884
	IM	0.5781	0.6095	0.5613	0.8510	0.6659	0.5143
	EF	0.5812	**0.6205**	**0.5701**	0.9038	**0.7142**	**0.5278**
	LF	**0.5813**	0.6187	0.5650	**0.9046**	0.7070	0.5190

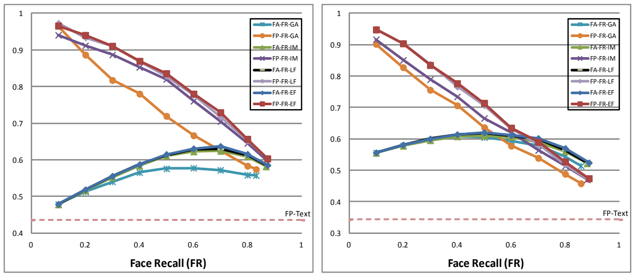

Figure 2: FA-FR and FP-FR curves of GA, IM, EF-IMGA and LF-IMGA on LYN (left) and WC (right).

4. CONCLUSION

We have proposed the IM, EF-IMGA and LF-IMGA methods for automatically associating faces appearing in images (or videos) with their names. The experiments conducted on both LYN and WC datasets basically validate our proposal, from which significant performance improvements are observed when compared with GA. The measurement of the similarity between heterogeneous face and name, though, is limited by the fact that faces "in the wild" are of high diversity. Thus, future work includes the incorporation of more contextual cues to further strengthen the calculation of the name-face similarity. We are also interested in generalizing the methods to dealing with less famous people.

5. ACKNOWLEDGMENTS

The research is supported by the National Natural Science Foundation of China (#61303175, #61272290, #61473030).

6. REFERENCES

[1] T. L. Berg, A. C. Berg, et al, Name and face in the news, IEEE CVPR, pp. 848-854, 2004.

[2] M. Guillaumin, T. Mensink, J. Verbeek, Face recognition from caption-based supervision. Int. J. Comput. Vis., 96(1): 64-82, 2012.

[3] J. Bu, B. Xu, C. Wu, et al, Unsupervised face-name association via commute distance. ACM Multimedia, pp. 219-228, 2012.

[4] S. Satoh, Y. Nakamura, T. Kanade, Name-it: naming and detecting faces in news videos. IEEE MultiMedia, 6(1): 22-35, 1999.

[5] M. Everingham, J. Sivic, A. Zisserman, Hello! my name is ... buffy -- automatic naming of characters in TV videos, In BMVC, pp. 889-908, 2006.

[6] M. Xu, X. Yuan, J. Shen, S. Yan, Cast2Face: character identification in movie with actor-character correspondence. ACM Multimedia, pp. 831-834, 2010.

[7] Y. Zhang, C. Xu, H. Lu, Character identification in feature-length films using global face-name matching. IEEE Trans. on Multimedia, 11(7):1276-1288, 2009.

[8] D. Ozkan, P. Duygulu, Interesting faces: a graph-based approach for finding people in news. Pattern Recognition, 43(5): 1717-1735, 2010.

[9] M. Zhao, J. Yagnik et al, Large-scale learning and recognition of faces in Web videos. IEEE FGR, pp. 1-7, 2008.

[10] Z. Chen, C.W. Ngo, J. Cao, W. Zhang, Community as a connector: associating faces with celebrity names in Web videos. ACM Multimedia, pp. 809-812, 2012.

[11] Z. Chen, C.W. Ngo, W. Zhang, J. Cao, Y. G. Jiang, Name-face association in Web videos: a large-scale dataset, baselines, and open issues. J. Comput. Sci. Technol., 29(5): 785-798, 2014.