Automatic Name-Face Alignment to Enable Cross-Media News Retrieval

Yuejie Zhang*, Wei Wu*, Yang Li*, Cheng Jin*, Xiangyang Xue*, Jianping Fan+
*School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, China
+Department of Computer Science, The University of North Carolina at Charlotte, USA
*{yjzhang, 10210240122, 11210240052, jc, xyxue}@fudan.edu.cn, +jfan@uncc.edu

Abstract

A new algorithm is developed in this paper to support automatic name-face alignment for achieving more accurate cross-media news retrieval. We focus on extracting valuable information from large amounts of news images and their captions, where multi-level image-caption pairs are constructed for characterizing both significant names with higher salience and their cohesion with human faces extracted from news images. To remedy the issue of lacking enough related information for rare name, Web mining is introduced to acquire the extra multimodal information. We also emphasize on an optimization mechanism by our Improved Self-Adaptive Simulated Annealing Genetic Algorithm to verify the feasibility of alignment combinations. Our experiments have obtained very positive results.

1 Introduction

With the explosive growth of multimodal news available both online and offline, how to integrate multimodal information sources to achieve more accurate cross-media news retrieval becomes an important research focus [Datta *et al.*, 2008]. Usually multimodal news is exhibited with the form of captioned news images, which mostly describe stories about people [Liu *et al.*, 2008]. Thus enabling automatic name-face alignment has become a critical issue for supporting cross-media news retrieval. However, because of the semantic gap [Fan *et al.*, 2012], there may exist huge uncertainty on the correspondence relationships among names (in captions) and faces (in images) [Deschacht *et al.*, 2007], [Yang *et al.*, 2009], e.g., the relationship between names and faces in a news is many-to-many rather than one-to-one.

To achieve automatic name-face alignment, three inter-related issues should be addressed simultaneously: 1) multimodal analysis of cross-media news to identify better correspondences between names and faces; 2) discovery of the missing information for rare name; and 3) multimodal optimization to achieve more effective alignments for all possible name-face pairs. To address the first issue, it is very important to develop robust algorithms that are able to achieve more accurate name-face alignment. To address the second issue, it is very interesting to leverage large-scale Web news for missing information prediction. To address the third issue, it is critical to develop a new optimization algorithm with high accuracy rate but low computational cost.

Based on these observations above, a novel scheme is developed in this paper for facilitating automatic name-face alignment to enable more effective cross-media news retrieval. Our scheme significantly differs from other earlier work in: a) Salient names and cohesive faces are extracted from cross-media news. Such names might have higher correspondence possibility with the faces in news images, and such faces might have higher coherence with the salient names. b) Web news is integrated for discovering missing information for rare name. c) An efficient measurement and optimization mechanism is established to verify the feasibility of our automatic alignment algorithm. d) A new algorithm is developed to achieve more precise characterization of the correlations between names and faces. Such cross-media alignment is treated as a problem of bi-media semantic mapping, and modeled as a correlation distribution over semantic representations of names and faces. Our experiments on a large number of public data from *Yahoo! News* have obtained very positive results.

2 Related Work

Alignment of names and faces in cross-media news is not a novel task [Satoh *et al.*, 1997&1999], [Berg *et al.*, 2004&2005]. Earlier research considered acquiring the relevant faces based on the original text query over image captions, and ranking or filtering the returned images by a face detector. However, the irrelevant or incorrect results may be produced by utilizing the simple matching between a query name and captions. Most face recognition approaches are only applied to the controlled setting and limited data collections [Mensink *et al.*, 2008], [Bozorgtabar *et al.*, 2011]. Thus the increasing research interest focuses on combining textual and visual information to support precise alignment.

In recent years, there is some related research work by using textual information that accompanies the image. Yang *et al.* [2004] proposed an approach for finding the specific persons in broadcast news videos by exploring various clues such as names in the transcript, face, anchor scenes and the timing pattern between names and people. Everingham *et al.*

[2006], [2009] investigated the problem of automatically labeling appearances of the names in TV or film, and demonstrated that high precision can be achieved by combining multiple sources of information. Ozkan *et al.* [2006] proposed a graph-based method to retrieve the correct faces of a queried person using both text and visual appearances. Le *et al.* [2007], [2008] introduced an unsupervised method for face annotation by mining Web, which aimed to retrieve relevant faces of one person by learning the visual consistency among results retrieved from text-correlation-based search engines. It's necessary to mention the most interesting research work developed by Berg *et al.* [2004], [2005], [2007], Guillaumin *et al.* [2008], [2012], and Pham *et al.* [2010], [2011]. Berg *et al.* proposed a method for association of names to faces using more realistic dataset. Guillaumin *et al.* considered two scenarios of naming people in databases of news photos with captions, that is, finding faces of a single query person and assigning names to all faces. Pham *et al.* reported their experiments on aligning names and faces as found in the images and captions of online news websites.

Unfortunately, all these existing approaches have not provided good solutions for the following important issues:

(1) Analysis and Measure for Name Salience and Name-Face Cohesion in Deep Level – Most existing alignment methods focus on exploiting all the inter-linkage information between each name and each face in a same captioned news image, which leads to a serious combinatorial problem for possible name-face pairs. Without sufficient analysis of name salience and measure for name-face cohesion, some names and faces may form the noisy information and cause the insignificant name-face matching judgement.

(2) Discovering Extra Multimodal Information for Rare Name – Most existing work concerns finding faces of a specific name underlying such a precondition that the relevant face set consists of a large group of highly similar faces for the original name query. With the increasing growth of Web news, it appears that such Web information can mitigate the lack of relevant faces for rare name. According to our best knowledge, no existing research has made full use of Web news to achieve more accurate alignment for rare name.

(3) Multimodal Alignment Optimization for Name-Face Pairs – Finding the best matching between names and faces for all the pairs is very difficult and complex, and may become a *NP*-hard problem. Probability-statistical models can be used for computing the alignment likelihood, but have the serious problem of getting stuck in a local maximum. From the view of combinatorial optimization, it's a very significant way to set a global objective function, the local constraint conditions and an integer programming model.

3 Name Salience Identification

In a news caption, not all names are equally important. Thus it's necessary to evaluate every name in the same caption and judge which names may possibly co-occur with the faces in the original image-caption pair. Here, "*salience*" is defined as a measurement for the important degree of each name in a caption. A pattern for computing such a score is constructed based on the in-depth multi-level analysis of caption.

In the initial name list, the names are enumerated without any ranking. Since the syntactic structure is indicative of various information distribution in a caption, the relative rank of names can be confirmed by means of the roles of names and the relationships among names in the syntactic parse tree.

(1) Syntactic Parse Tree Depth (SPTD) – This factor indicates the minimum depth for all the names of a certain name clustering in the parse tree for a caption. If a name has a shallower depth, it will have the higher importance. Given a caption with *N* names, each name is associated with its corresponding name clustering NC_i and NC_{ij} represents the j^{th} name in NC_i. The depth value for each NC_i can be defined as:

$$SPTD(NC_i) = \min_{j=1}^{SNC(NC_i)} \{SPT\_Depth(NC_{ij})\} \qquad (1)$$

where $SNC(NC_i)$ denotes the size of NC_i, i.e., the number of names with the co-reference relationship contained in NC_i, which can also be understood as the occurrence frequency of the same names with different forms in the given caption; and $SPT\_Depth(NC_{ij})$ is the depth of NC_{ij} in the parse tree.

(2) Syntactic Parse Tree Traversal Order (SPTTO) – This factor indicates the minimum breadth-first traversal order for all the names of a certain name clustering in the parse tree for a caption. It's proposed based on such a fact that among the nodes in the same level, a node with the higher priority of the traversal order means its higher relative importance. The traversal order value for each NC_i can be defined as:

$$SPTTO(NC_i) = \min_{j=1}^{SNC(NC_i)} \{SPT\_BFT - Order(NC_{ij})\} \qquad (2)$$

where $SPT\_BFT\text{-}Order(NC_{ij})$ is the breadth-first traversal order of NC_{ij} in the syntactic parse tree.

In most cases the higher frequency for a specific name in a caption indicates its higher importance. It's very necessary to take such frequency into account as a key reference factor. Given *N* different names in a caption, in which each name corresponds to a name clustering NC_i, i=1, …, N, the initial rank value, *Relative Salience* (*RS*), can be computed as:

$$RS(NC_i) = \alpha * \frac{SNC(NC_i)}{\sum_{j=1}^{N} SNC(NC_j)} + \beta * \frac{\sum_{j=1}^{N} SPTD(NC_j) - SPTD(NC_i)}{(N-1) * \sum_{j=1}^{N} SPTD(NC_j)} \qquad (3)$$

$$+ \gamma * \frac{\sum_{j=1}^{N} SPTTO(NC_j) - SPTTO(NC_i)}{(N-1) * \sum_{j=1}^{N} SPTTO(NC_j)}, \sum_{i=1}^{N} RS(NC_i) = 1$$

where α, β and γ are the relative importance coefficients, $\alpha+\beta+\gamma=1$. Maybe there is an extreme situation that only one name is involved in the caption, it's insignificant to evaluate the relative salience for such a unique name, $RS(NC_1)=1$.

It's rather crude to assume that every name in the caption appears in the image. Thus it's very significant to make the further refinement to filter the names with the extremely lower possibility to be pictured as the detected faces. A visual survey for name-face pairs with the association linkage learned that the number of available faces is generally less than or equal to the number of detected names in a pair. We make the simplifying assumption that all the names with the initial rank values below a certain threshold may not be pictured as the selected faces. Given *F* faces and *N* names (*F*<=*N*) in the initial rank list for a pair, *F*' selected faces (*F*'<=*F*) are retained after face filtering and *N* names are refined to *N*' names (*F*'<=*N*'<=*N*) according to the visual exhibition information. Under such refinement, multiple names with the relatively higher importance are chosen to interpret the selected faces in the image, i.e., name denoising.

4 Name-Face Cohesion Measure

The general assumption for name-face alignment is that the appearances of the faces for a person/name are more similar than those for different persons/names. Thus it's very useful to establish a cohesion measurement to evaluate the intrinsic association degree among all the faces in the whole database.

Firstly, the similarity between two faces is evaluated based on the common neighbors among k-nearest neighbors for every face. Given two faces F_i and F_j in the local face set FS_N_m related to a real name N_m, if F_i is close to F_j and they are both close to FS_N_m, F_i and F_j are close with higher confidence for the mutual density since their similarity is determined based on all the faces in FS_N_m. Thus the similarity between a pair of faces can be defined as:

$$Sim(F_i, F_j) = \frac{|KNS(F_i, FS\_N_m, k) \cap KNS(F_j, FS\_N_m, k)|}{k} \quad (4)$$

where $KNS(F_i, FS\_N_m, k)$ and $KNS(F_j, FS\_N_m, k)$ denote the k-nearest neighbors when F_i or F_j positions in FS_N_m; and k is a dynamically varied value and changes according to the size of the local face set related to the current name.

Secondly, the *Local Density Score* (*LDS*) is utilized to measure the density for each face in the local face set. The larger the density is, the more relevant the face and the name associated with the face set are. The *LDS* for a face F_i can be described as the average similarity between this face and its k-nearest faces in the same face set. The higher *LDS* indicates the higher connectivity between F_i and its k-nearest faces, that is, F_i is more relevant with N_m associated with FS_N_m.

$$LDS(F_i, FS\_N_m, k) = \frac{\sum_{F_j \in KNS(F_i, FS\_N_m, k)} Sim(F_i, F_j)}{k} \quad (5)$$

Thirdly, the *Local Cohesion Degree* (*LCD*) is especially proposed to measure the mutual density among all the faces in the face set related to a particular name under the current global alignment manner, and defined as:

$$LCD(FS\_N_m, k) = \sum_{F_i \in FS\_N_m} LDS(F_i, FS\_N_m, k) \quad (6)$$

If a face has the higher *LDS* value in the related face set, this face will have the higher density with the other faces in the same face set and then will be more relevant with the name corresponding to the face set. Therefore, as the sum of the *LDS* values for all the faces in a local face set, *LCD* can be well representative of the mutual density among all the faces in this face set. The larger the *LCD* is, the higher the mutual density among all the faces in the face set will be. The higher *LCD* can be regarded as a good indication of the higher local cohesion for the face set, that is, the face set is more relevant with the assigned name. Under an arbitrary global alignment manner, the sum of the *LCD* values for all the face sets can reflect the whole global cohesion for the complete face set related to all the names. Thus the larger sum of the *LCD* values for the face sets embodies the higher global cohesion for the complete face set, and then further expresses the higher global relevance between the whole face set and the name set under the current global alignment manner.

5 Web Mining for Rare Name

The name-face alignment can also be regarded as a restricted face naming or retrieval problem. The general assumption is that the number of faces corresponding to the query name is relatively large. However, some rare names may lead to such a fact that the number of their faces is very small. The "*rare name*" here does not mean the real single name, but refers to the rare person that occurs only several times or even one time in the whole database. Thus for the alignment evaluation of *rare name*, it's necessary to establish a discovery mechanism to supplement more available multimodal information.

As the vast knowledge base, Web has become the best resource to create a reference annotated face image set for rare name. Face images about a specific person can be automatically retrieved by using Web image search engines. Based on Web mining, a certain number of relevant face images and their annotations can be discovered simultaneously, and become the valuable enrichment for rare name. All the name expression forms for a rare name are collected in a rare name clustering. Each form in the clustering is taken as an original name query and submitted to Web image search engines. Thus a list of relevant face images and their annotations can both be acquired. Each returned face image will be processed through the face detection and filtering, and the available and meaningful face images are kept.

Given a rare name $Name_0$, all of its face images can be ranked according to the salience value of $Name_0$ in each face image. The *Top-R* face images are maintained as the complementary face information to $Name_0$. R is a generalized threshold for the complementary face image selection, and does not have a fixed setting. Because there may be different numbers of face images returned for different rare names, R means a limited proportion of the face images preferred, and varies with the size of the returned face image set.

6 Name-Face Alignment via ISSAGA

We consider such an alignment task as an optimization problem for searching the optimal name-face matching in each pair. The Standard Genetic Algorithm (SGA) can obtain the global optimal search using genetic operations, but there are some disadvantages such as "*prematurity*" and the poor search power for local optimal solutions [Wang *et al*., 2005]. However, the Simulated Annealing Algorithm (SAA) has the stronger ability for local optimization [Andresen *et al*., 2008]. Thus the Improved Self-Adaptive Simulated Annealing Genetic Algorithm (ISSAGA), which integrates the advantages of SGA and SAA and a self-adaptive probability adjustment strategy, is introduced to make a better solution.

6.1 Multimodal Name-Face Alignment

Suppose P image-caption pairs include F real faces (i.e., *Not Null Face*) and N real names (i.e., *Not Null Name*), PS denotes the Image-Caption Pair Set for P pairs; FS denotes the Face Set for F faces; NS denotes the Name Set for N names; W_FP_{ij} denotes whether the face F_j belongs to the pair P_i, $P_i \in PS$, $F_j \in FS$, i=1, …, $|PS|$, j=1, …, $|FS|$, which is equal to:

$$W\_FP_{ij} = \begin{cases} 1, & \text{if the face } F_j \text{ is in the pair } P_i \\ 0, & \text{otherwise} \end{cases}$$

W_NP_{ik} denotes whether the name N_k belongs to P_i, $P_i \in PS$, $N_k \in NS$, i=1, …, $|PS|$, k=1, …, $|NS|$, which is equal to:

$$W\_NP_{ik} = \begin{cases} 1, & \textit{if the name } N_k \textit{ is in the pair } P_i \\ 0, & \textit{otherwise} \end{cases}$$

FP_i denotes the face set for P_i, $FP_i=\{F_j|W\_FP_{ij}=1, F_j\in FS\}$, $P_i\in PS$, i=1, …, $|PS|$; NP_i denotes the name set for P_i, $NP_i=\{N_k|W\_NP_{ik}=1, N_k\in NS\}$, $P_i\in PS$, i=1, …, $|PS|$; W_FN_{kj} denotes whether F_j is assigned to N_k, $F_j\in FS$, $N_k\in NS$, j=1, …, $|FS|$, k=1, …, $|NS|$, which is equal to:

$$W\_FN_{kj} = \begin{cases} 1, & \textit{if the face } F_j \textit{ is assigned to the name } N_k \\ 0, & \textit{otherwise} \end{cases}$$

FS_N_k denotes the total associated face set for N_k based on the global assignment of names and faces, $N_k\in NS$, k=1, …, $|NS|$; $LCD(FS\_N_m, k)$ denotes the mutual density among all the faces in the face set FS_N_m related to the name N_m under a certain global alignment manner, $N_m\in NS$, m=1, …, $|NS|$. Thus our mathematical model can be formalized as:

$$\max \sum\nolimits_{N_m\in NS} \frac{LCD(FS\_N_m, k)}{|FS\_N_m|} = \max \sum\nolimits_{N_m\in NS} \frac{\sum_{F_i\in FS\_N_m} LDS(F_i, FS\_N_m, k)}{|FS\_N_m|} \quad (7)$$

s.t., 1) $\bigcup\limits_{1\le i\le|PS|} FP_i = FS$, $\bigcup\limits_{1\le i\le|PS|} NP_i = NS$. That is, the union set of the face set included in each pair is the universal face set of the whole database, and the union set of the name set included in each pair is the universal name set. 2) $FP_i\cap FP_j=\Phi$, $i\neq j$, i, j=1, …, $|PS|$. That is, there is no intersection set between the face sets of two pairs. 3) $|FP_i|=|NP_i|$, i=1, …, $|PS|$. That is, to establish more efficient coding information for the optimization algorithm, we form a constraint that the numbers of faces and names in an arbitrary pair must be same. Null faces or names can be used to supplement the less ones. 4) $\sum_{k=1}^{|NP_i|} W\_FN_{kj} \le 1$, $F_j\in FP_i$, $N_k\in NP_i$, i=1, …, $|PS|$, j=1, …, $|FP_i|$. That is, in a pair each real face can only be assigned to at most a real name in the same pair. 5) $\sum_{j=1}^{|FP_i|} W\_FN_{kj} \le 1$, $F_j\in FP_i$, $N_k\in NP_i$, i=1, …, $|PS|$, k=1, …, $|NP_i|$. That is, in a pair each real name can only be assigned to at most a real face in the same pair. 6) $\sum_{F_j\in FP_i, N_l\notin NP_i} W\_FN_{lj} = 0$, i=1, …, $|PS|$, j=1, …, $|FP_i|$, l=1, …, $|NS|$. That is, the face that belongs to a pair can only be assigned to the name in the same pair.

6.2 ISSAGA-based Multimodal Optimization

a. Encoding of Chromosome

The names from all the pairs are positioned in a fixed order, and the faces from all the pairs are ranked in segments and then grouped into a chromosome. Each chromosome corresponds to a solution. A repeatable natural number encoding pattern is adopted to design the gene g_{ij} of the chromosome C, $C=\{g_{ij}\}$, i=1, …, $|PS|$, j=1, …, $|FP_i|$, where i is the pair number; j is the face number for the pair P_i; and g_{ij} denotes the number for the face F_j in P_i. C can be expanded as $\{g_{11}, …, g_{1|FP1|}, …, g_{i1}, …, g_{i|FPi|}, …, g_{|PS|1}, …, g_{|PS||FP_{|PS|}|}\}$, where $\{g_{i1}, …, g_{i|FPi|}\}$ is a segment and each segment keeps the relatively independent relationship with the other segments.

b. Initial Population

The initial population $P(t)$ with L chromosomes is produced by a stochastic method. The stochastic order in each segment is used to generate an initial chromosome C, and then the other L-1 different chromosomes are achieved from C. According to the objective function value, the current best chromosome can be regarded as the initial optimal solution.

c. Self-Adaptive Selection-Reproduction Operator

According to our mathematical model, the objective function for the l^{th} chromosome in a population can be constructed as:

$$OF(C_l) = \sum\nolimits_{N_m\in NS} \frac{LCD(FS\_N_m, k)}{|FS\_N_m|} \quad (8)$$

To limit the quantity of chromosomes that dissatisfy the constraints in the new population, we use a Multiple Roulette Wheel method and compute the selection probability as:

$$P_S(C_l) = \frac{f'(C_l)}{\sum_{k=1}^{M} f'(C_k)}, \quad l=1,\cdots,M \quad (9)$$

where M is the number of chromosomes in the current population, and $f'(\ )$ is obtained through the self-adaptive transformation based on the original fitness function $f(\ )$ to avoid "*prematurity*" and "*slow convergence speed*".

$$f(C_l) = OF(C_l), \quad f'(C_l) = a * f(C_l) + \frac{e - e^{g/g_{\max}}}{e + e^{g/g_{\max}}} * (f_{\max} - f_{\min}) \quad (10)$$

where f_{max} and f_{min} are the maximum and minimum fitness values for the current population; g is the number of generations under the current genetic manner; g_{max} is the maximum number of genetic generations; and a is a constant parameter.

d. Self-Adaptive Simulated Annealing Crossover Operator

Considering in most cases there are two or three faces in a news image and the maximum order number for the faces in a segment of a chromosome is also two or three, thus a One-Point crossover method is adopted. Two special strategies of the self-adaptive crossover probability (P_C) computation and the simulated annealing operation are introduced.

$$P_C(C_i, C_j) = \begin{cases} P_{C1} - \dfrac{(P_{C1} - P_{C2})(\max(f(C_i), f(C_j)) - f_{avg})}{f_{\max} - f_{avg}}, & \max(f(C_i), f(C_j)) \ge f_{avg} \\ P_{C1}, & \max(f(C_i), f(C_j)) < f_{avg} \end{cases} \quad (11)$$

where f_{avg} is the average fitness value in the current population; P_{C1} and P_{C2} are two predefined parameters. The simulated annealing operation is utilized to judge whether the inferior solution should be accepted to substitute the previous chromosome when facing with an inferior solution and generate a probability (P_A) for accepting the inferior solution.

$$P_A(C_l) = \frac{1}{1 + e^{(f(C_l') - f(C_l)) / (T_0 * \delta^g)}} \quad (12)$$

where $f(C_l')$ denotes the fitness value for the new chromosome C_l' generated after the crossover; T_0 is the initial temperature for the simulated annealing operation; and δ is a preset ratio coefficient for the temperature reduction.

e. Self-Adaptive Simulated Annealing Mutation Operator

The mutation operator cannot be executed among segments but inside every segment by utilizing an Exchange Mutation method. Meanwhile, two special strategies of the self-adaptive mutation probability (P_M) computation and the simulated annealing operation are also introduced.

$$P_M(C_l) = \begin{cases} P_{M1} - \dfrac{(P_{M1} - P_{M2})(f_{\max} - f(C_l))}{f_{\max} - f_{avg}}, & f(C_l) \ge f_{avg} \\ P_{M1}, & f(C_l) < f_{avg} \end{cases} \quad (13)$$

where P_{M1} and P_{M2} are two predefined parameters. Given a chromosome $C_k=\{S_1, …, S_i, …, S_N\}$, $N=|PS|$, S_i is the i^{th} segment. Taking a segment as a unit, the 2-exchange mutation is accomplished among the segments of C_k by $P_M(C_k)$.

7 Experiment and Analysis

7.1 Dataset and Evaluation Metrics

Our dataset is established based on *Labeled Yahoo! News Data* constructed by Berg *et al*. [2005] and further developed by Guillaumin *et al*. [2008]. It consists of 20,071 captioned news images, and all the image-caption pairs contain 31,147 face images that belong to 5,873 different persons/names. To evaluate the effectiveness of our algorithm, the ground truth associations are considered to measure the *benchmark metrics* for name-face aligned pairs. *Correct Label Rate* (*CLR*) is defined as the percentage of correct pairs generated in the ground truth links. Our evaluation can not only consider the association links between real names and faces, but also concern the assignments to null names or null faces.

7.2 Experiment on Cross-Media Alignment

Our cross-media alignment model is created by integrating Name Salience Ranking (NSR), Name-Face Cohesion Measure (NFCM), Web-based Multimodal Information Mining (WMIM) for rare name and the Improved Self-Adaptive Simulated Annealing Genetic Algorithm of ISSAGA. To investigate the effect of each part on the whole alignment performance, we introduce three evaluation patterns: 1) *Baseline*(*NFCM*) – based on the basic preprocessing with *No NSR* for captions and *NFCM* for images; 2) *Baseline*(*NFCM*)+*NSR* – based on the baseline model with *NFCM*, *NSR* is integrated; 3) *Baseline*(*NFCM*)+*NSR*+*WMIM* – by fusing the baseline model with *NFCM* and *NSR*, *WMIM* is added. These patterns are tested in two alignment optimization manners of general GA and ISSAGA, which aims to explore the advantages of our proposed ISSAGA.

To verify the consistency of our alignment model in different mapping cases, the performance evaluation focuses on two assignment schemes. One contains the specific mappings of *Real Face->Null Name* (*RF->NN*) and *Real Name->Null Face* (*RN->NF*) except the general mapping of *Real Face<->Real Name* (*RF<->RN*), and another only involves *RF<->RN*. The experimental results are shown in Table 1.

Assignment Scheme	Evaluation Pattern	GA	ISSAGA
		***CLR* (%)**	***CLR* (%)**
RF<->RN ***RF->NN*** ***RN->NF***	***Baseline*(*NFCM*) [*No NSR*]**	52.12	53.49
	Baseline*(*NFCM*)+*NSR	58.81	60.76
	Baseline*(*NFCM*)+*NSR*+*WMIM	**61.71**	**63.16**
RF<->RN	***Baseline*(*NFCM*) [*No NSR*]**	54.31	56.78
	Baseline*(*NFCM*)+*NSR	61.17	62.47
	Baseline*(*NFCM*)+*NSR*+*WMIM	**63.21**	**65.77**

Table 1. The experimental results on our whole dataset.

It can be seen from Table 1 that for the name-face linking in our whole dataset, we can still obtain the best *CLR* value of 65.77% in the same evaluation pattern of fusing *Baseline* with *NFCM*, *NSR*, *WMIM* and *ISSAGA* and excluding the mappings related to *Null Name* and *Null Face*. In comparison with the baseline model, the alignment performance could be promoted to a great degree by successively adding *NSR* and *WMIM* under our *ISSAGA*-based optimization. Compared the results for different assignment schemes, we can observe that when only considering the alignment between real names and faces, the alignment performance increases and becomes pretty good. Although when evaluating the links with null names and faces the alignment performance maybe slightly influenced by such extra uncertainty, the *CLR* can still reach a relatively high value. Through the comparison between the results based on the traditional GA and ISSAGA, it can be found that our ISSAGA is obviously superior to the traditional GA and more suitable for the alignment optimization.

In addition, the convergence curves of the best solutions found by the traditional GA and our ISSAGA are shown in Figure 1. The traditional GA improves the solutions very fast within 60 iterations, and reaches a plateau after that. In contrast, our ISSAGA keeps constantly improving the solutions with more iterations, gives much better solutions, and has a stronger ability of escaping from the local optima. It can be concluded that our ISSAGA provides better solutions within a satisfactory amount of time and is very useful to contribute better solutions for such a special optimization problem.

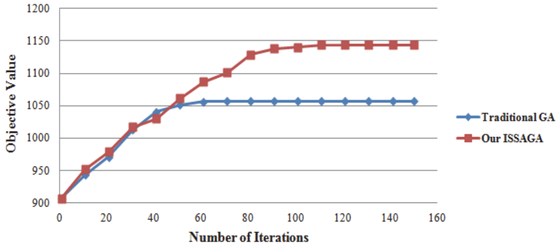

Figure 1. The convergence curves of the traditional GA and our ISSAGA.

7.3 Experiment on Cross-Media News Retrieval

To further explore the applicability of our alignment model in cross-media news retrieval, we particularly select 16 names with the larger number of faces in the whole dataset as the query topics and carry out the retrieval runs based on the original dataset and the dataset with the alignment processing respectively. The *Precision* (*P*), *Recall* (*R*) and *F-measure* (*F*) values for each name query are shown in Table 2. It can be seen that the best run is based on the dataset with alignment processing, and its results exceed those by another run based on the original dataset without any alignment processing dramatically. For all the name queries, their *F* values under the run with alignment processing are obviously higher than those under another run, and the highest increase can reach to 20%. By adopting our alignment model for the original dataset, the whole retrieval performance for the selected name queries has gained the significant improvement.

Name Query	Captioned News Image Dataset for Testing					
	The Original Dataset			The Dataset with Alignment Processing		
	Baseline+NSR			*NSR+NFCM+WMIM+ISSAGA*		
	***P* (%)**	***R* (%)**	***F* (%)**	***P* (%)**	***R* (%)**	***F* (%)**
George W. Bush	55.30	95.72	**70.11**	75.44	87.69	**81.11**
Collin Powell	57.38	72.92	**64.23**	75.00	69.08	**71.92**
Tony Blair	62.59	80.45	**70.41**	77.49	78.12	**77.80**
Donald Rumsfeld	70.77	76.72	**73.63**	87.22	75.57	**80.98**
Gerhard Schroeder	56.88	80.52	**66.67**	77.83	74.46	**76.11**
Ariel Sharon	77.55	22.75	**35.18**	97.44	22.75	**36.89**
Hugo Chavez	47.26	90.32	**62.05**	86.61	88.71	**87.65**
Junichiro Koizumi	55.78	73.21	**63.32**	73.11	77.68	**75.32**
Jean Chretien	62.67	86.24	**72.59**	80.73	80.73	**80.73**
John Ashcroft	55.82	83.49	**66.91**	76.85	76.15	**76.50**
Jacques Chirac	48.25	73.40	**58.22**	63.72	76.60	**69.57**
Hans Blix	47.10	81.11	**59.59**	73.96	78.89	**76.34**
Vladimir Putin	42.76	75.58	**54.62**	51.85	65.12	**57.73**
Silvio Berlusconi	55.88	71.25	**62.64**	66.27	68.75	**67.48**
Arnold Schwarzenegger	60.00	88.73	**71.59**	95.45	88.73	**91.97**
John Negroponte	72.22	60.00	**65.55**	86.36	58.46	**69.72**

Table 2. The experimental results for cross-media retrieval.

7.4 Comparison with Existing Approaches

To give full exhibition to the superiority of our alignment model, we have also performed a comparison between our approach and the other classical ones in recent years. Three approaches developed by Berg *et al*. [2005], Guillaumin *et al*. [2008] and Pham *et al*. [2010] are analogous with ours, and then we accomplished them on the same dataset. The experimental results are presented in Table 3, which reflect the difference of power between these four approaches.

Approach	Assignment Scheme	Evaluation Pattern	*CLR* (%)
Berg *et al*.'s [2005] (*Berg*)	***RF<->RN RF->NN RN->NF***	***Berg* [*No NSR*]**	**49.58**
		Berg+NSR	53.07
		Berg+NSR+WMIM	**57.12**
	RF<->RN	***Berg* [*No NSR*]**	**52.71**
		Berg+NSR	56.45
		Berg+NSR+WMIM	**59.93**
Guillaumin *et al*.'s [2008] (*Guill*)	***RF<->RN RF->NN RN->NF***	***Guill* [*No NSR*]**	**52.14**
		Guill+NSR	57.87
		Guill+NSR+WMIM	**60.97**
	RF<->RN	***Guill* [*No NSR*]**	**54.41**
		Guill+NSR	59.95
		Guill+NSR+WMIM	**62.43**
Pham *et al*.'s [2010] (*Pham*)	***RF<->RN RF->NN RN->NF***	***Pham* [*No NSR*]**	**52.69**
		Pham+NSR	56.13
		Pham+NSR+WMIM	**61.57**
	RF<->RN	***Pham* [*No NSR*]**	**55.61**
		Pham+NSR	58.98
		Pham+NSR+WMIM	**62.73**
Our Approach with *ISSAGA*	***RF<->RN RF->NN RN->NF***	***Baseline(NFCM)* [*No NSR*]**	**53.49**
		Baseline(NFCM)+NSR	60.76
		Baseline(NFCM)+NSR+WMIM	**63.16**
	RF<->RN	***Baseline(NFCM)* [*No NSR*]**	**56.78**
		Baseline(NFCM)+NSR	62.47
		Baseline(NFCM)+NSR+WMIM	**65.77**

Table 3. The comparison between our and the other existing methods.

It can be found from Table 3 that for the name-face linking on our whole dataset by Berg/Guiilaumin/Pham *et al*.'s approaches, we can obtain the best *CLR* values of 52.71%, 54.41% and 55.61% in the evaluation patterns of *Berg*/*Guill*/*Pham* [*No NSR*] and excluding the mappings related to *Null Name* and *Null Face* respectively. The main reason is that when only considering the alignments between real faces and real names, all these three approaches can reduce the combination ambiguities between names and faces in the same image-caption pair to a certain degree and then the relatively better *CLR* values can be obtained. Compared the results of *Berg*/*Guill*/*Pham* [*No NSR*] and our baseline model with *NFCM*, we can find the best *CLR* value of 56.78% appears in the results of our model, which is obviously higher than those *CLR* values of 52.71%, 54.41% and 55.61% for *Berg*/*Guill*/*Pham* [*No NSR*] respectively. This implies that both of our name-face cohesion measure mechanism *NFCM* and combination optimization algorithm *ISSAGA* are feasible for facilitating more effective name-face alignment evaluation and optimization. Furthermore, through fusing our *NSR* and *WMIM* with Berg/Guillaumin/Pham *et al*.'s approaches in turn, the better performance can be acquired, in which the best *CLR* values are considerably improved to 59.93%, 62.43% and 62.73% respectively. This confirms the prominent roles of our *NSR* and *WMIM* in automatic name-face alignment once again. Compared with the improved Berg/Guillaumin/Pham *et al*.'s approaches that integrate with *NSR* and *WMIM*, our alignment model with *NFCM*, *NSR*, *WMIM* and *ISSAGA* can still present the better performance and the best *CLR* value of 65.77% differs greatly from those of Berg/Guillaumin/Pham *et al*.'s. This indicates that our approach is really superior to Berg/Guillaumin/Pham *et al*.'s, and also further confirms that our alignment model with *NFCM*, *NSR*, *WMIM* and *ISSAGA* is exactly a better way for determining name-face associations and can support such a special combinatorial optimization problem more effectively.

7.5 Analysis and Discussion

Through the analysis for the aligned name-face linkages with failure, it can be found: 1) The alignment acquisition is associated with the preprocessing for image-caption pairs, especially for caption text. It's easier for image and caption preprocessing to produce some errors or missing detections for faces and names. Such wrong or undetected information will seriously affect the whole alignment performance. 2) The other names that co-occur with a certain name in the same caption maybe helpful for facilitating this name's related face finding and naming. The same case is for faces in images. Such name or face co-occurrence can be utilized to further improve the alignment measurement. 3) For multimodal information from Web mining, instead of directly using the feedback from image search engines, it's helpful to exploit a relevance evaluation manner to select more appropriate complementary information for rare name. (4) It is a novel way to solve name-face alignment from the view of combinatorial optimization. Although our ISSAGA have obtained the satisfactory performance on large-scale dataset, making more highly specialized consideration about genetic operations will be more beneficial to acquiring more feasible solutions. 5) Some news captions involve extremely rare name. With very limited information supplement from Web, such news lacks enough multimodal information for alignment processing. This may be a more stubborn problem.

8 Conclusions and Future Work

A new algorithm is introduced in this paper to support more precise automatic name-face alignment for cross-media news retrieval. The multi-level analysis of image-caption pairs is established for characterizing the meaningful names with higher salience and the cohesion between names and faces. To remedy the issue of rare name, the extra multimodal information is acquired through Web mining. ISSAGA is introduced to solve the alignment optimization better. Thus a novel model is developed by integrating all these aspects to effectively exploit name-face correlations. Our future work will focus on making our system available online, so that more Internet users can benefit from our research.

Acknowledgments

This work is supported by National Natural Science Fund of China (No. 61170095; No. 61272285), National Science & Technology Pillar Program of China (No. 2012BAH59F04), 973 Program of China (No. 2010CB327906), Shanghai Outstanding Academic Leader Program (No. 12XD1400900), and Shanghai Leading Academic Discipline Project (No. B114). Cheng Jin is the corresponding author.

References

[Andresen *et al*., 2008] M. Andresen, H. Bräsel, J. Tusch, M. Mörig, F. Werner, and P. Willenius. Simulated annealing and genetic algorithms for minimizing mean flow time in an open shop. *Mathematical and Computer Modelling*, 48(7-8):1279-1293, 2008.

[Berg *et al*., 2005] T.L. Berg, A.C. Berg, J. Edwards, and D.A. Forsyth. Who's in the picture. *Advances in Neural Information Processing Systems 17*, 37-144, 2005.

[Berg *et al*., 2007] T.L. Berg, A.C. Berg, J. Edwards, and M. Maire. Names and faces. Technical Report, U.C. at Berkeley, 2007.

[Berg *et al*., 2004] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. The, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proceedings of CVPR 2004*, 2:848-854, 2004.

[Bozorgtabar *et al*., 2011] B. Bozorgtabar and G.A. Rezai Rad. A genetic programming - PCA hybrid face recognition algorithm. *Journal of Signal and Information Processing*, 2:170-174, 2011.

[Datta *et al*., 2008] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* (*CSUR*), 40(2), Article 5, 2008.

[Deschacht *et al*., 2007] K. Deschacht and M.F. Moens, Text analysis for automatic image annotation. In *Proceedings of ACL2007*, 1000-1007, 2007.

[Everingham *et al*., 2006] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name Is … Buffy - Automatic naming of characters in TV video. In *Proceedings of BMVC 2006*, 889-908, 2006.

[Everingham *et al*., 2009] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5):545-559, 2009.

[Fan *et al*., 2012] J.P. Fan, X.F. He, N. Zhou, J.Y. Peng, and R. Jain. Quantitative characterization of semantic gaps for learning complexity estimation and inference model selection. *IEEE Transactions on Multimedia*, 14(5):1414-1428, 2012.

[Guillaumin *et al*., 2008] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In *Proceedings of CVPR 2008*, 1-8, 2008.

[Guillaumin *et al*., 2012] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 96(1):64-82, 2012.

[Le *et al*., 2008] D.D. Le and S. Satoh. Unsupervised face annotation by mining the Web. In *Proceedings of ICDM 2008*, 383-392, 2008.

[Le *et al*., 2007] D.D. Le, S. Satoh, M.E. Houle, and D.P.T. Nguyen. Finding important people in large news video databases using multimodal and clustering analysis. In *Proceedings of ICDEW 2007*, 127-136, 2007.

[Liu *et al*., 2008] C. Liu, S. Jiang, and Q. Huang. Naming faces in broadcast news video by image Google. In *Proceedings of MM 2008*, 717-720, 2008.

[Mensink *et al*., 2008] T. Mensink and J. Verbeek. Improving People Search using Query expansions: How friends help to find people. In *Proceedings of ECCV 2008*, 86-99, 2008.

[Ozkan *et al*., 2006] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photo. In *Proceedings of CVPR 2006*, 1477-1482, 2006.

[Pham *et al*., 2010] P.T. Pham, M.F. Moens, and T. Tuytelaars. Cross-media alignment of names and faces. *IEEE Transactions on Multimedia*, 12(1):13-27, 2010.

[Pham *et al*., 2011] P.T. Pham, T. Tuytelaars, and M.F. Moens. Naming people in news videos with label propagation. *IEEE Multimedia*, 18(3): 44-55, 2011.

[Satoh *et al*., 1997] S. Satoh and T. Kanade. Name-it: Association of face and name in video. In *Proceedings of CVPR 1997*, 368-373, 1997.

[Satoh *et al*., 1999] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and Detecting Faces in News Videos. *IEEE Multimedia*, 6(1):22-35, 1999.

[Wang *et al*., 2005] Z.G. Wang, M. Rahman., and Y.S. Wong. Optimization of multi-pass milling using parallel genetic algorithm and parallel genetic simulated annealing. *International Journal of Machine Tools and Manufacture*, 45(15):1726-1734, 2005.

[Yang *et al*., 2004] J. Yang, M.Y. Chen, and A.G. Hauptmann. Finding person x: Correlating names with visual appearances. In *Proceedings of CIVR 2004*, 270-278, 2004.

[Yang *et al*., 2009] Y. Yang, D. Xu, F.P. Nie, J.B. Luo, and Y.T. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of MM 2009*, 175-184, 2009.