Cast Search via Two-Stream Label Propagation

Jhih-Ciang Wu[1,2], Bing-Jhang Lin[1,2], Bing-Yuan Zeng[2], Li-Chen Fu[2], Chiou-Shann Fuh[2], Tyng-Luh Liu[1]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]National Taiwan University, Taiwan

Abstract

We address the problem of Cast Search by Portraits (CSP) where a facial portrait of a cast member is provided to retrieve from a given video clip those frames containing the query target. The underlying CSP formulation is related to the task of person re-identification. However, CSP is more challenging in that the provided query image is only a portrait of a certain cast member, and the instances of the target in the candidate video could have a very different visual appearance. Such drastic visual variations are not common in addressing the problem of person re-id. We propose a two-stream network architecture for tackling the CSP challenge and also participate in the public CSP competition[1]. The overall outcome in the competition is promising and worth further effort to improve our proposed model.

1. Introduction

Cast Search by Portraits (CSP) is a practical problem of image retrieval. The goal is to uncover image frames of a given video clip which are *relevant* to the query portrait. Notice that the CSP task is similar to the problem of person re-identification (re-id). However, the latter is more specific in that the person re-id task typically seeks to match people appearing in different views of a multi-camera surveillance system within a short duration [9, 14, 13, 16]. The pivotal difference between CSP and person re-id is the visual disparity of the query target and candidates (see Figure 2). The instances (both the target and candidates) of each particular identity are *similar* in the task of person re-id, while particularly emphasizing the aspects of visual appearances of age, clothing, scene, *etc*. But such similarities evolve into visual disparity in the cast search, which makes CSP more challenging than person re-id.

Motivated by that each cast portrait is depicted with a face image of high quality, we use an effective facial recognition model to extract discriminative features such that strong visual links between a cast member and the candidates can be effectively explored. To this end, the label information is propagated through the face and the ID-discriminative embedding (IDE) [15] visual links in the Progressive Propagation via Competitive Consensus (PPCC) scheme [8]. However, extra attention needs to be paid to those candidates that visual information from the face area is less reliable due to, say, poor image quality, and occlusions. We consider a two-stream architecture for tackling this task. The first stream focuses on facial information that we employ hands-on face detectors and design a Convolutional Neural Network (CNN) model for feature extraction. It is possible that some candidates may have incomplete face information due to occlusions blurs, noises, or pose changes. Such a concern is common in solving the task of person re-id. We thus design the second complementary stream that adopts a conventional person re-id model and aims to deal with the frames without facial information from face detectors. Finally, we obtain the prediction via coupling the two streams with a label propagation algorithm. For more details, please refer to Section 3.

Figure 1. CSP (top) vs. Person re-id (bottom): While both tasks aim to uncover a subset of candidates matching to the given query, the aspects of visual similarity to be explored are rather distinct.

2. Related work

Person search The person search problem, e.g., [11] typically aims to seek the ground-truth images in the gallery with respect to the query. The task is relevant to person re-id but more challenging since the bounding box information is unavailable.

[1]Challenge website: http://wider-challenge.org/2019.html

Figure 2. Illustration of visual variation in CSP. Left: a cast portrait; Right: a possible candidate image.

Person re-identification While the most related task to CSP is person re-id, a direct application of person re-id for solving CSP may lead to ineffective inference [8]. The most contrasting part with respect to person re-id is that CSP uses only a *portrait* to carry out the query task.

3. Our method

Inspired by PPCC, the label information can be propagated through the face and the IDE visual links. Figure 3 illustrates our framework which comprises a two-stream network. The two-stream network focuses on building the face and the IDE visual links respectively. However, some candidates with poor image quality have less "face information" mentioned in Section 1 such that it requires extra iterations in the algorithm of label propagation. In our work, the fundamental concept is to better initialize the visual links for propagating the label information effectively.

3.1. Face model

The first stream focuses on facial information that we adopt hands-on face detectors and use a CNN model for feature extraction. We take both a cast portrait and candidates as input. Each cast portrait is depicted with a face image of high quality, so we adopt the face visual links as the initial relationship between the portrait and candidates.

For face detection, we respectively adopt the Multi-task Cascaded Convolutional Networks (MTCNN) [12] for the cast and the Face Alignment Network (FAN) [1] for candidates. Owing to the nature of a cast portrait, we can easily detect the face of the cast portrait by a tiny model based on cascaded regression. On the contrary, it is advantageous to use a deeper model, often boosted with the attention mechanism, for candidates due to the variability of candidates. There may be multiple human faces in a single candidate image after the face detection. We select a *main* face as the candidate. Figure 4 indicates our selection scheme. We choose a face to represent a candidate by position (1) and the intersection (2) of the face region and the whole candidate image. Observe that the face of the main actor in a candidate is usually in the middle of the image. The constant c in (1) controls the position of the whole candidate image. We formulate a scoring function as expressed in (3).

$$S_h = 1 - \frac{|m_f - m_w \times c|}{m_w \times c} \tag{1}$$

where m_f and m_w are the respective median positions of the detected faces and the whole candidate image along the horizontal axis. c is a constant that adjusts the position of the candidate image. An analogous formulation for S_v along the vertical axis is adopted.

$$S_a = \frac{a_f}{a_w} \tag{2}$$

where a_f and a_w are the area of each detected face and the whole candidate image respectively.

$$S_{total} = S_h \times S_v \times S_a \tag{3}$$

After performing face detection and selection, the process continues to carry out feature extraction. We use ResNet50 [6] and SENet [7] as the backbone which are pretrained on MS-Celeb [5] and fine-tuned by VGGFace2 [2]. We use the commonly metric, cosine similarity for calculating the affinity by the extracted features.

3.2. Re-ranking face visual links

The initial relationship between the cast portrait and candidates is defined by the face model. It yields mediocre performance due to the less face information in candidates. We notice that these candidates are similar to the other candidates in top-k ranking even though they are not similar to the cast portrait. Thus the initial relationship can be refined by the re-ranking algorithm [17]. The Mahalanobis distance [3] is used for measuring the similarity of the appearance feature, and the k-reciprocal nearest neighbors of each candidate is defined by the Jaccard metric in [17]. The cosine similarity is used to be the measurement metric for the appearance feature. It is better to convert the affinity matrix A with its components $a_{ij} \in [-1, 1]$ into a distance matrix. The affinity matrix A is transformed by a Gaussian kernel in (4) which maps the affinity matrix to a better representation. We take the distance matrix D as input to the re-ranking algorithm for refinement. After the re-ranking processing, the distance matrix maps to the original range by (5).

$$D = e^{-\frac{(1-A-2)^2}{2\sigma^2}} \tag{4}$$

where σ indicates the bandwidth of the Gaussian kernel; A is the original affinity matrix; and D represents the transformed distance matrix.

$$M_{re} = 1 - rank(D) \tag{5}$$

where M_{re} is the processed affinity matrix, and the $rank(\cdot)$ denotes the re-ranking algorithm [17].

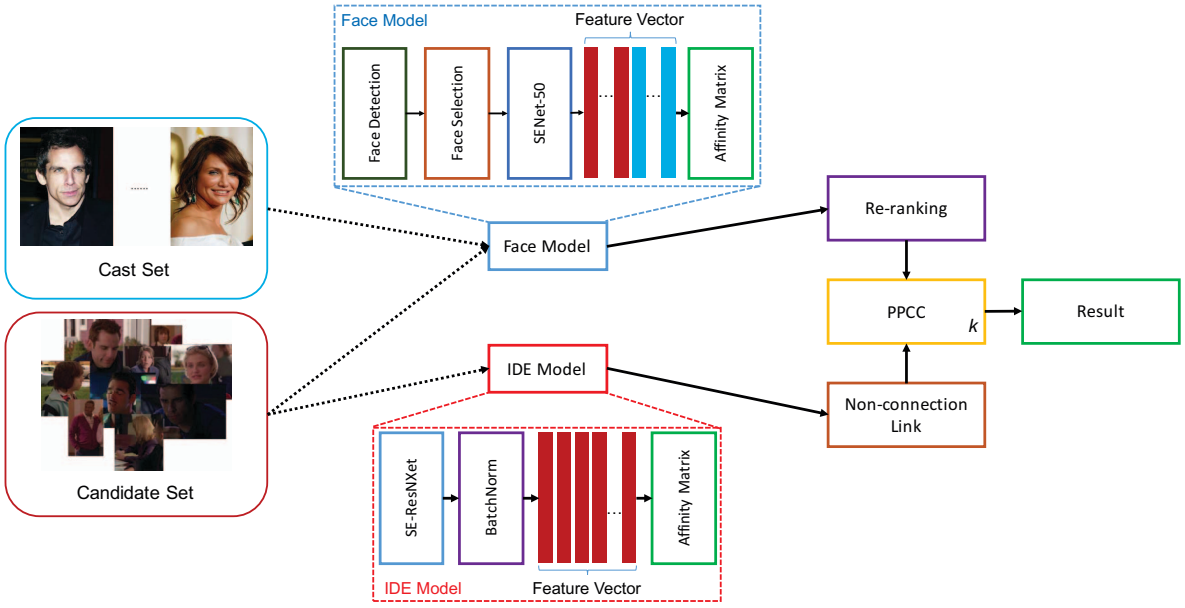

Figure 3. Overview of our framework. We design a two-stream architecture to tackle the CSP problem. We obtain the prediction via coupling the two streams with a label propagation algorithm.

3.3. IDE model

The IDE model focuses on building the visual links between candidates. Although some candidates have poor quality images for face information due to the occlusions, blurs, noises, or pose changes, *etc*., the dominant visual feature is the clothing. This challenge is the same as the person re-id problem. We follow a number of training tricks [10] for training the IDE model. There are three objective functions for the IDE model: softmax, triplet, and center loss. During the training stage, the anchor of the triplet loss is the candidate set since the IDE model is focusing on the relationship between candidates. Besides, there is a visual disparity of cast and candidates as we mention before. It is hard to build the visual link. The triplet loss and center loss focus on clustering the same label data in a group instead. For making ranking concept more effective, we propose the ranking loss (6) following the scoring metric of mean Average Precision (mAP) in this task.

$$loss_{rank} = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{1}{M}\sum_{j=1}^{M}[L_i = L_j]P(j)) \qquad (6)$$

where N is the number of the anchors, and M shows number of samples (positive and negative) from the gallery set. $[\cdot]$ is the Iverson Bracket function. In inference stage, the IDE affinity matrix is also calculated by cosine similarity through the extracted IDE features.

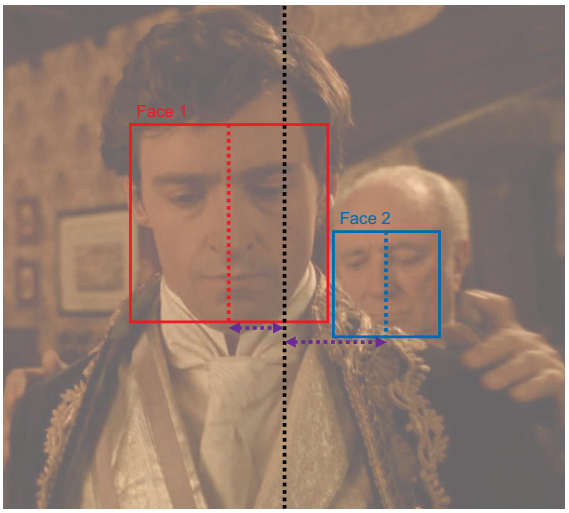

Figure 4. An example from our face selection scheme. There are two faces detected by FAN and denoted as red and blue bounding boxes, respectively. The dotted line divides the image or the bounding box in half vertically. The purple dotted arrows show the distance between the dotted lines.

3.4. IDE non-connective links

It is possible that there could be more than one candidate in a singe movie frame (see Figure 5). On the other hand, it is unlikely that an actor appears in the same scene more than once. Therefore, a non-connective value can be set to those candidate pairs in the IDE affinity matrix. The similarity values of these candidate links should be minimum.

Figure 5. An example of the non-connective situation that there are two candidates in a movie frame. The two red bounding boxes indicate two different candidates.

3.5. Propagating label information

Our method builds a graph accounting for the cast and candidates and iterates by propagating the label information through the visual and temporal links in PPCC. Still it takes some iterations to pass the information due to possible poor image quality of candidates. We refine two kinds of affinity matrices for getting better initial visual links. After the refinement, the PPCC is applied for achieving the prediction result. The iteration of PPCC is set as $k = 2$ in our framework. The first iteration aims to pass the label information through the stronger face affinity. The second uses the IDE affinity for building the relationship with those candidates with less face information. Both two initial methods encode the better pairwise relationship for the cast and candidates. Nevertheless, it is not necessary to do more iterations for propagating the label information in our experiment.

4. Experiments

The proposed model is evaluated mainly by the competition of the WIDER Cast Search by Portrait Challenge 2019. We report two conventional evaluation metrics: top-k accuracy and mean Average Precision (mAP).

4.1. Dataset

The dataset is an extension of CSM [8]. It comes from 630 movies and split into subsets of 250, 180, and 200 clips for training, validation, and testing respectively. The cast of each movie is collected from IMDb or TMDb. The candidates are captured from the key frames of each movie with manually-labeled person bounding boxes.

4.2. Comparison with different visual links

We analyze the influences of different visual links shown in Table 1. There are three comparisons: the IDE visual links only, the face visual links only, and both the IDE and the face visual links. Our result shows that the face is much more important than the IDE while building the connection between the cast and candidates. The joint consideration yields more satisfactory results under our proposed setting.

Methods	mAP
IDE	0.0881
face	0.6429
IDE + face	**0.8055**

Table 1. Performance comparison on the different visual links.

Methods	mAP	top-1	top-5
baseline	0.7654	0.9431	0.9610
baseline + N	0.7669	0.9431	0.9610
baseline + R	0.8008	0.9411	0.9583
baseline + N + R	0.8055	0.9418	0.9590
baseline + N + R + ensemble	**0.8281**	**0.9563**	**0.9689**

Table 2. Comparison on the validation set of the competition.

Team	mAP	Team	mAP
Jiaoda Poets	0.7671	MCC_USTC	0.8730
SAT_ICT	0.7466	VIPL_ICT	0.8515
MCC_USTC	0.7402	IIS_CVL	0.8137

Table 3. Leaderboard of WIDER Cast Search by Portrait. Left: 2018; Right: 2019. Our work is indicated as team IIS_CVL which achieves the 3rd place in the 2019 competition.

4.3. Ablation study

We compare the performance of our method with the PPCC. The backbone for the adopted face extractor is SENet pre-trained on MS-Celeb and fine-tuned by VGG Face2. The IDE model is pre-trained on ImageNet [4] and uses the softmax, triplet, center, and ranking loss for training. We compare three scenarios: (I) Using non-connective links in the IDE affinity matrix (N); (II) Re-ranking the face affinity matrix (R); (III) Using both initial methods. These three scenarios use the same affinity matrix. The temperature of softmax in PPCC is set $t = 0.05$ for all three options. We set $\sigma = 0.8$ for the bandwidth of Gaussian kernel.

Table 2 shows our ablation study. Using non-connective links just slightly improves the baseline since the IDE visual links are not the major connection between the cast and candidates. It shows that re-ranking has significant improvement. We obtain the best result in each single model by using both initial methods. Besides, we adopt the ensemble strategy and get the highest score for our model.

Acknowledgement This work was supported in part by the MOST, Taiwan under Grant 108-2634-F-001-007. We are also grateful to the *National Center for High-performance Computing* for providing computational resources and facilities.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 2

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 2

[3] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[5] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2016. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2

[8] Qingqiu Huang, Wentao Liu, and Dahua Lin. Person search in videos with one portrait through visual and temporal links. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 425–441, 2018. 1, 2, 4

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 1

[10] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[11] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1

[12] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2

[13] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 1

[14] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 1

[15] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[16] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 1

[17] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3652–3661, 2017. 2