

10-2010

Cast2Face: Character identification in movie with actor-character correspondence

Mengdi XU

National University of Singapore

Xiaotong YUAN

National University of Singapore

Jialie SHEN


Singapore Management University, jlshen@smu.edu.sg

Shuicheng YAN

National University of Singapore

DOI: <https://doi.org/10.1145/1873951.1874090>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

Citation

XU, Mengdi; YUAN, Xiaotong; SHEN, Jialie; and YAN, Shuicheng. Cast2Face: Character identification in movie with actor-character correspondence. (2010). *MM '10: Proceedings of the 18th ACM International Conference on Multimedia: Firenze, Italy, October 25-29, 2010*. 831-834. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/651

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Cast2Face: Character Identification in Movie with Actor-Character Correspondence

Mengdi Xu ¹, Xiaotong Yuan ¹, Jialie Shen ², Shuicheng Yan ¹

¹ Department of Electrical and Computer Engineering, National University of Singapore, Singapore

² School of Information Systems, Singapore Management University, Singapore
{g0900224,eleyuanx, eleyans}@nus.edu.sg, {jshen}@smu.edu.sg

ABSTRACT

We investigate the problem of automatically identifying characters in a movie with the supervision of actor-character name correspondence provided by the movie cast. Our proposed framework, namely Cast2Face, is featured by: (i) we restrict the names to assign within the set of character names in the cast; (ii) for each character, by using the corresponding actor's name as a key word, we retrieve from Google image search a group of face images to form the gallery set; and (iii) the probe face tracks in the movie are then identified as one of the actors by robust multi-task joint sparse representation and classification method. The assigned actor name of a face track is then mapped to the character name based on the cast again. In addition to face naming, we further apply the proposed method to spotlights summarization of a particular actor in his/her movies. Empirical evaluations on several feature-length movies demonstrate the satisfying performance of our method.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems-Evaluation/methodology; I.2.10 [Vision and Scene Understanding]: Video analysis; I.5.4 [Applications]: Computer Vision

General Terms

Experimentation, Performance

Keywords

Character Identification, Cast Analysis, Video Summarization, Face Recognition, Multi-Task Learning

1. INTRODUCTION

The problem of character identification concerns identifying the faces of the characters in a movie and label them with their corresponding names. In a feature-length movie,

the characters are often the most important contents to be indexed, and thus character identification becomes a critical step on film semantic analysis. As has been noted by previous studies [1][2][3] that although very intuitive to humans, automatic character identification is still tremendously challenging due to: 1) the lack and ambiguity of available annotations; and 2) many factors other than identity (e.g., pose, light and expression) influence the way a face appears in a frame.

In this work, we present a novel cast analysis and image retrieval based approach for automatically naming the faces of the characters in a movie. One basic motivation is that the cast of a film is always available and the internet provides a vast of information on the actors. Typically, a film cast contains the names of actors, characters and their (one-one) correspondence. We propose to do a matching between the faces detected from the movie and a gallery set of face images searched from web for the actors in the cast. The assigned actor name of a face is then mapped to the character name by the actor-character correspondence. This work is different from the state-of-the-art name-to-face methods [1][2][12] where subtitle and/or scripts are required. Based on the results of character identification, a further application to generate spotlights summarization and digestion of a particular actor in many of his/her movies is also presented.

1.1 Related Work

The task of associating faces with names in a movie or TV program is typically accomplished by combining multiple sources of information, e.g. image, video and text, under less or even no manual intervention. Extensive research efforts have been concentrated on this task. Arandjelovic and Zisserman [1] used face image as a query to retrieve particular characters. Affine warping and illumination correcting were utilized to alleviate the effects of pose and illumination variations. In [5], a multi-cue approach combining facial features and speaker voice models was proposed for major cast detection. However, these approaches cannot automatically assign real names to the characters. To handle this, Everingham *et al.* [2] proposed to employ readily available textual sources, the film script and subtitle, for text video alignment and thus obtained certain annotated face exemplars. The rest of the faces were then classified into these exemplars for identification. Their approach was also followed by Laptev *et al.* [4] for human action annotation. However, in their approach [2], the subtitle text and time-stamps were extracted by OCR, which required extra computation cost on spelling error correction and text verification. Moreover,

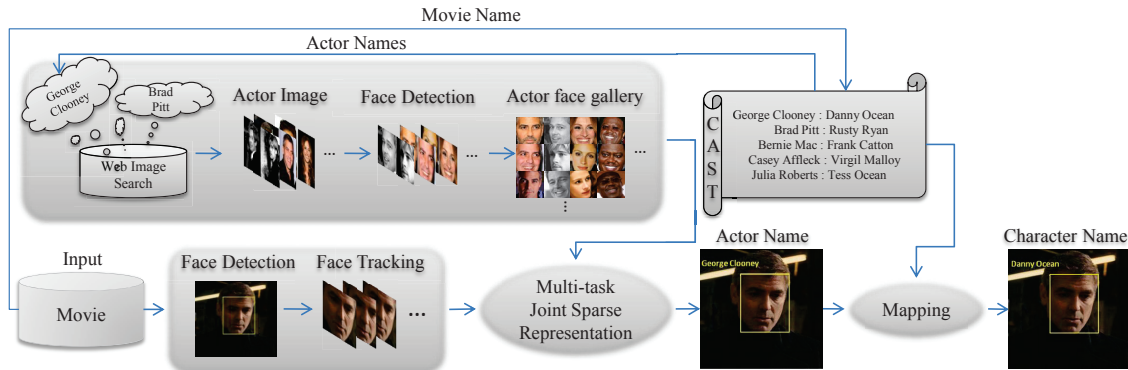


Figure 1: Scheme illustration of Cast-to-Face method.

for some movies, the scripts can not be found easily or the scripts may be quite different from subtitles. In this work, we aim to develop a name-to-face method without costly textual analysis.

1.2 Outline of Our Method

The Cast2Face method we describe is a novel framework for labeling the faces of the characters in a movie with cast. Our method comprises three components: 1) gallery face set collection with cast analysis and web image search; 2) probe face tracks extraction and description; and 3) face tracks identification with robust multi-task joint sparse representation and classification. Compared with previous studies on name-to-face, the main contributions of this paper include:

1. To the best of our knowledge, Cast2Face is the first work combing character identification with cast analysis and web image retrieval.
2. A robust multi-task joint sparse representation method is developed to classify each face track without training on possibly contaminated gallery set.
3. We design a novel application of our method to automatically generate the spotlights summarization of a particular actor in many of his/her movies.

Figure 1 depicts the working mechanism of our proposed Cast2Face method.

2. THE CAST-TO-FACE SYSTEM

2.1 Cast based Web Image Search and Gallery Set Generation

In order to associate names with characters detected in a movie, we use the movie cast list which is easily available on the web or at the end of the movie. We restrict the names to assign within the set of character names appeared in the cast. For each character, by using the corresponding actor’s name as a key word, we retrieve from Google image search a set of images. We observe that the top hundreds of the returned images belong to the actor with high precision. We then employ a frontal face cascade detector [9] included in OpenCV2.0¹ to detect and crop faces from the downloaded images. In this way, the gallery set is established, and then

¹<http://sourceforge.net/projects/opencvlibrary>

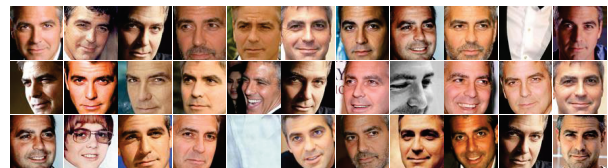


Figure 2: Some exemplar faces of the star actor George Clooney in the gallery set generated by web image search and face detection.

used for labeling the faces of the characters extracted from the movie. Taking the movie “Ocean’s Twelve” as an example, some gallery face images for the key actor, George Clooney, are shown in Figure 2. Note that a few incorrect faces are inevitably introduced in the gallery set due to image retrieval and face detection errors. As we shall see in later on experiments that our face identification method is quite robust to such noises contained in the gallery set.

2.2 Probe Face Tracks Extraction and Description

In this step, we again use the frontal face cascade detector [9] to detect faces appeared in each frame of the movie. A typical movie may contain tens of thousands of detected faces. However, these faces merely arise from a few hundred “tracks” of a particular character. Therefore it is plausible to discover the correspondences between faces to reduce the volume of data need to be processed. Furthermore, stronger appearance models can be built for each character since a face track provides multiple examples of the character’s appearance. To obtain face tracks, a robust foreground correspondence tracker [11] is applied for each shot. Here shot changes are automatically detected using color histogram difference between consecutive frames. The short tracks which are often introduced by false positive detections are discarded.

We construct the representation of a face by a part-based descriptor extracted around local facial features [2]. Here we first use a generative model [1] to locate nine facial key-points in the detected face region, including the left and right corners of each eye, the two nostrils and the trip of the nose and the left and right corners of the mouth. We then extract the 128-dim SIFT [6] descriptor from each key-point and concatenate them to form the face descriptor with

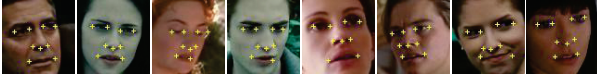


Figure 3: Examples of detected face with facial feature points.

dimensionality 1,152. Figure 3 illustrates some selected faces with facial feature points marked.

2.3 Multi-Task Joint Sparse Representation and Classification

Given a set of retrieved gallery face images and the extracted probe face tracks, we present in this section a simple yet efficient algorithm for face track identification.

Each unlabeled face track is, nevertheless, simply represented as a set of image feature vectors extracted from all images in the track. One simple method for identification, as conducted in [2], is to directly calculate the feature distance between a probe face track and the labeled exemplar faces, and then assign probe face track to the nearest neighborhood. Another feasible method is to classify each image in the track independently via, e.g., sparse representation classification [10], and then assign the face track to the subject that achieves the highest frequency.

In this work, by viewing the identification of each image in a probe face track as a task, the face track identification can be naturally casted to a multi-task face recognition problem. This motivates us to apply the multi-task joint sparse representation model [7] to face track classification. The key advantage of multi-task learning lies in that it can efficiently make use of complementary information contained in different sub-tasks.

Suppose we have a set of exemplar faces with M subjects. Denote $X = [X_1, \dots, X_M]$ as the feature matrix where $X_m \in \mathbb{R}^{d \times p_m}$ is associated with the m -th subject. Here d is the dimensionality of features and $p = \sum_{m=1}^M p_m$ is the total number of samples. Given a probe face track as an ensemble of L images $\{y^l\}_{l=1, \dots, L}$, $y^l \in \mathbb{R}^d$, we consider a supervised L -task linear representation problem as follows:

$$y^l = \sum_{m=1}^M X_m w_m^l + \varepsilon^l, l = 1, \dots, L, \quad (1)$$

where $w_m^l \in \mathbb{R}^{p_m}$ is a reconstruction coefficient vector associated with the m -th subject, and ε^l is the residual term. Denote $w^l = [(w_1^l)^T, \dots, (w_M^l)^T]^T$ the representation coefficients for the probe image feature y^l , and $w_m = [w_m^1, \dots, w_m^L]$ the representation coefficients from the m -th subject across different images. Furthermore, we denote $W = [w_m^l]$. Therefore, our proposed multi-task joint sparse representation model is formulated as the solution to the following problem of multi-task least square regressions with $\ell_{1,2}$ mixed-norm regularization:

$$\min_W F(W) = \frac{1}{2} \sum_{l=1}^L \left\| y^l - \sum_{m=1}^M X_m w_m^l \right\|_2^2 + \lambda \sum_{m=1}^M \|w_m\|_2. \quad (2)$$

Here, we use the popular optimization method of Accelerated Proximal Gradient (APG) [8] to solve the Eqn. (2) with fast convergence rate guaranteed.

When the optimal $\hat{W} = [\hat{w}_m^l]$ are obtained, a probe image y^l can be approximated as $\hat{y}^l = X_m \hat{w}_m^l$. For classification,

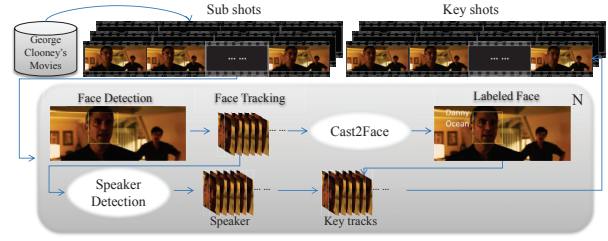


Figure 4: Framework of Actor-Specific Spotlights summarization.

the decision is ruled in favor of the subject with the lowest total reconstruction error accumulated over all the L tasks:

$$m^* = \arg \min_m \sum_{l=1}^L \left\| y^l - X_m \hat{w}_m^l \right\|_2^2. \quad (3)$$

We call model (2) along with classification rule (3) as the multi-task joint sparse representation and classification (*MTJSRC*) in this paper.

3. APPLICATION: ACTOR-SPECIFIC SPOTLIGHTS SUMMARIZATION

Based on the results of character identification, there are many applications, such as character-specific movie retrieval, personalized video summarization, intelligent playback and video semantic mining, etc. Here we apply Cast2Face to actor-specific spotlights summarization, on which users can input the actor names to search and digest the film content.

We first divide the movie into several sub-shots with scene change detection. Each shot is about 1~2 minutes duration. After the identification of all the detected face tracks in these shots, we rank the tracks associated with a particular actor according to the reconstruction error calculated in rule (3). The video shots containing the top 10 tracks are then taken as the candidate spotlight videos. We further restrict that the actor should be speaking in the summarized video. The speaker is identified using visual information, i.e., finding face detections with significant lip motion [2]. We then combine the obtained key shots together as a digested movie. Figure 4 illustrates the working scheme of the proposed actor-specific movie summarization method.

4. EXPERIMENTS

The Cast2Face method along with its application in spotlights summarization is empirically evaluated on several feature-length movies. All the movie casts are obtained from the Internet Movie Database (IMDB)².

4.1 Performance of Character Identification via MTJSRC

As a quantitative study of Cast2Face, we evaluate in this experiment the accuracy of our proposed MTJSRC method for character identification. We report the corresponding results on three films “Ocean’s Twelve”(2004), “Titanic”(1997) and “Twilight”(2008). The sizes of the constructed gallery sets for some selected actors are listed in Table 1. Two baseline methods are employed for comparison: 1) the nearest neighbor(NN) classifier used in [2] which directly calculates

²<http://www.imdb.com/>

Table 1: The quantitative results on the evaluation of Cast2Face method.

Movie Name	Actor Name	Gallery Set Size	Probe Face Tracks	Total Faces	Accuracy(%)		
					NN	SR	MTJSRC
Ocean’s Twelve	George Clooney	90	27	1,477	63.4	63.1	73.3
	Julia Roberts	94	28	1,677	30.0	41.9	82.1
	Matt Damon	96	32	1,310	72.6	73.8	87.5
	Brad Pitt	90	13	522	43.8	72.6	92.3
Titanic	Leonardo DiCaprio	43	19	429	37.0	69.6	68.4
	Kate Winslet	57	41	1,015	62.4	72.2	82.9
	Billy Zane	47	18	386	57.7	80.0	88.8
Twilight	Kristen Stewart	96	122	3,321	82.4	84.5	89.3
	Robert Pattinson	92	98	2,612	70.8	82.8	89.8
	Taylor Lautner	47	6	86	16.6	53.3	50.0

Table 2: Quantitative Results on Spotlights Summarization for actor *George Clooney*.

Movie Name	Total Tracks	Positive	True Positive
Ocean’s Eleven	232	23	19
Ocean’s Twelve	228	17	15
Up In The Air	274	40	38

the feature distance between a probe face track and the labeled exemplar faces, and then assign probe face track to the nearest neighborhood; and ii) the sparse representation (SR) classifier [10] which classifies each image in the track independently and then assign the face track to the subject that most frequently occurs in this track. As aforementioned, there always exist a few incorrect faces in the gallery set, thus training based methods, e.g., SVM and Subspace analysis, are not applicable in our setting. In contrast, our multi-task linear representation based method is quite robust to the condemnation since the joint representation ability of noise images are low comparing to those “good” samples.

The evaluation results are listed in Table 1, from which we can see that MTJSRC significantly outperforms both baselines for 8 out of the 10 testing actors. For computational cost, Cast2Face method is training free and the most expensive calculation lies in the testing phase where a multi-task regression problem (see Eqn. (2)) is optimized. In our experiment, the adopted APG algorithm converges at roughly 10~20 rounds of iterates. The average running time is 0.31s per probe face track. The parameter λ in Eqn. (2) is set to 0.1 throughout our experiment.

4.2 Actor-Specific Spotlights Summarization

In this experiment, we apply Cast2Face to spotlights summarization and evaluate the performance. We build a gallery set containing face images of 21 actors from three films “Ocean’s Eleven”(2001), “Ocean’s Twelve”(2004) and “Up In the Air”(2009). Taking actor George Clooney as an example, we aim to extract the key shots for him from these films. After the multi-task joint sparse representation and classification, we obtain a set of tracks identified as George Clooney, among which, the tracks including speakers are taken as the key tracks. Table 2 shows the tracks detection and identification results. The sub-shots including key track is called as key shot. By assembling these key shots we can get the final spotlights summarization for George Clooney. The result on this experiment and the results from previous experiment are online available in YouTube: <http://www.youtube.com/user/cast2face>.

5. CONCLUSION

Cast2Face is a novel cast and image retrieval based movie character identification method. We demonstrate that high precision can be achieved by combining multiple sources of information including the cast, web image and movie. Comparing to the subtitle and script based methods, one appealing aspect of our method is that it is textual analysis free. We also explored an application of our method for actor-specific spotlights summarization. Empirical evaluations on feature-length movies show the satisfying performance of Cast2Face method.

6. ACKNOWLEDGEMENT

This work is supported by National Research Foundation/Interactive Digital Media Program, under research Grant NRF2008IDMIDM004-029, Singapore.

7. REFERENCES

- [1] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *CVPR*, pages 860–867, 2005.
- [2] M. Everingham, J. Sivic, and A. Zisserman. ‘hello! my name is... buffy’ - automatic naming of characters in tv video. In *BMVC*, pages 889–908, 2006.
- [3] M. Everingham and A. Zisserman. Identifying individuals in video by combining generative and discriminative head models. In *ICCV*, pages 1103–1110, 2005.
- [4] I. Laptev, M. Marszalk, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
- [5] Z. Liu and Y. Wang. Major cast detection in video using both speaker and face information. *IEEE Transactions on Multimedia*, 9(1):89–101, 2007.
- [6] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [7] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Journal of Statistics and Computing*, 2009.
- [8] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal of Optimization*, 2008.
- [9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518. IEEE, 2001.
- [10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–226, Feb. 2009.
- [11] T. Yang, Q. Pan, J. Li, and S. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *CVPR*, pages 970–975. IEEE, 2005.
- [12] Y.-F. Zhang, C. Xu, H. Lu, and Y.-M. Huang. Character identification in feature-length films using global face-name matching. *IEEE Transactions on Multimedia*, 11(9):1276–1288, Nov. 2009.