Character Identification in TV-series via Non-local Cost Aggregation

Ching-Hui Chen
ching@umiacs.umd.edu
Rama Chellappa
rama@umiacs.umd.edu

Department of Electrical and Computer Engineering
University of Maryland
College Park, USA

Character identification is an important task for preparing the metadata for a TV-series, since several applications, such as video summarization [5], analysis of character interactions [7], and shot retrieval [3, 8], require knowing the identities of humans in the scenes. Nevertheless, character identification in TV-series remains a challenging task since the video is usually unconstrained and the human pose varies. Existing works use the names provided in the screenplay, speech identification [6], and attributes (e.g., gender) [1], to assist person identification. Nevertheless, several prior efforts [9, 10] based on face clustering and tracking are not suitable for consistently identifying the characters in a TV-series due to shot variations and the occlusion of faces. Since the human body is more perceivable even when the face is occluded, person tracking [2] can provide additional advantages for person identification.

Recently, non-local cost aggregation methods have been shown to yield good results in establishing dense stereo correspondence [4, 11]. This framework ensures that information is effectively utilized via non-local cost aggregation on minimum spanning tree. Motivated by these works, we propose a non-local identification framework to recognize the identity of each face track and person track such that identities in the venue are consistently reported. Extending the non-local framework to solve the identification problem in a TV-series is not straightforward. Unlike pixels with identical modalities which line up in a planar graph structure, face tracks and person tracks own different modalities in the timeline. Besides, contextual information (e.g., unique presence of a subject) should be utilized to improve the identification performance.

In this paper, we propose a unified approach for identifying face and person tracks in a TV-series. We construct the track nodes to multiplex the modalities of face and clothing feature from face and person track, respectively. Our method explores the video structure by constructing the minimum-distance spanning tree (MST) from the track nodes such that track nodes that are likely to have the same identity are adjacent to each other. A typical identification task assigns the identity such that the cost of each track node is minimized. By performing the non-local cost aggregation on MST, the identity assignment becomes more reliable by minimizing the aggregated cost, which allows the information from adjacent track nodes to be utilized as a whole. Furthermore, the unique presence of a subject in a venue is taken into account by minimizing the total aggregated cost of track nodes with temporal dependency. Experimental results on TV-series datasets demonstrate the effectiveness of the proposed method.

The block diagram of our approach is illustrated in Figure 1. First, the face and person tracks form track nodes if their bounding boxes co-occur with reasonable relative positions. Track nodes can be categorized into three types: Face-body, face-only, and body-only track node. Face-body track nodes consist of both modalities from face and person tracks (See Figure 2). Some track nodes only have a single modality, either from face or person track. Face-only track nodes appear when the human body cannot be detected. On the other hand, the body-only track nodes commonly appear when actors turn their bodies around. It is clear that face-body track nodes possess more information as compared to track nodes of a single modality. Track nodes with temporal dependency are then grouped into a knot. We construct the MST from the k-knot graph such that information can be conveyed across the track nodes in the video sequence. The cost of each track node is aggregated on the MST such that track nodes with similar appearance and temporal adjacency are more likely to have the same identity. The identification problem is thus cast as a cost minimization problem per knot such that the uniqueness constraint is incorporated. In the end, face and person tracks inherit the identity of the track node they are associated with.

We conduct experiments on two TV-series datasets. Experimental results confirm that the proposed method can effectively utilize the face and person tracks for character identification.

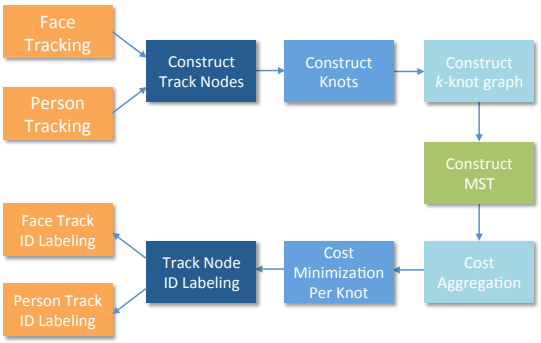

Figure 1: Block diagram of our proposed framework.

Figure 2: Face tracks (blue and green) and a person track (red) are merged into one track node.

[1] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[3] M. Everingham J. Sivic and A. Zisserman. Person spotting: Video shot retrieval for face sets. In *Image and Video Retrieval*, 2005.

[4] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] J. Sang and C. Xu. Character-based movie summarization. In *ACM Multimedia*, 2010.

[6] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" probabilistic person identification in TV-series. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Storygraphs: Visualizing character interactions as a timeline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *ACM International Conference on Multimedia Retrieval*, 2014.

[9] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[10] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[11] Q. Yang. A non-local cost aggregation method for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.