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

Abstract

We propose a non-local cost aggregation algorithm to recognize the identity of face and person tracks in a TV-series. In our approach, the fundamental element for identification is a track node, which is built on top of face and person tracks. Track nodes with temporal dependency are grouped into a knot. These knots then serve as the basic units in the construction of a k-knot graph for exploring the video structure. We build the minimum-distance spanning tree (MST) from the k-knot graph such that track nodes of similar appearance are adjacent to each other in MST. Non-local cost aggregation is performed on MST, which ensures information from face and person tracks is utilized as a whole to improve the identification performance. The identification task is performed by minimizing the cost of each knot, which takes into account the unique presence of a subject in a venue. Experimental results demonstrate the effectiveness of our method.

1 Introduction

Character identification is an important task for preparing the metadata for a TV-series, since several applications, such as video summarization [14], analysis of character interactions [16], and shot retrieval [9, 17], require knowing the identities of humans in the scenes. Nevertheless, character identification in a 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 [15], and attributes (e.g., gender) [6], to assist person identification. Furthermore, analyzing the text information in the subtitles has been used for person identification [13]. Nevertheless, several prior efforts [19, 20] 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 [8] 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 [11, 21]. This framework ensures that information is effectively utilized via non-local cost aggregation on the 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

Pages 119.1-119.11
DOI: https://dx.doi.org/10.5244/C.29.119

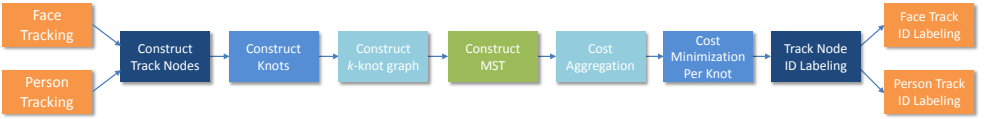

Figure 1: Block diagram of our proposed approach.

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 time-line. Besides, contextual information (e.g., unique presence of a subject) should be utilized to improve the identification performance.

Our work is closely related to [15] that models the character appearances as Markov Random Field (MRF) representations of face and person tracks. Additional information, such as alternating shots and speaker identification, is utilized into their model to improve the identification performance. MRF framework suggested in [15] relies on tracks with both face and clothing modalities, and transfers the face identification result to person tracks where faces cannot be authenticated due to occlusion or other reasons. However, this approach cannot guarantee that all information in face and person tracks is utilized as a whole since the procedure of pre-clustering of clothing appearance and post-assignment of the identity to person tracks based on clustering results are performed separately.

In this paper, we propose a unified approach for identifying the 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 possesses the capability to explore the video structure via 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 via 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.

2 The Proposed Framework

2.1 Overview of Our 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 (Section 2.2). Track nodes with temporal dependency are then grouped into a knot (Section 2.3). We construct the MST from the k-knot graph (Section 2.4) 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 (Section 2.5). The identification problem is thus cast as a cost minimization problem per knot such that the uniqueness constraint is incorporated (Section 2.6). In the end, face and person tracks inherit the identity of the track node they are associated with.

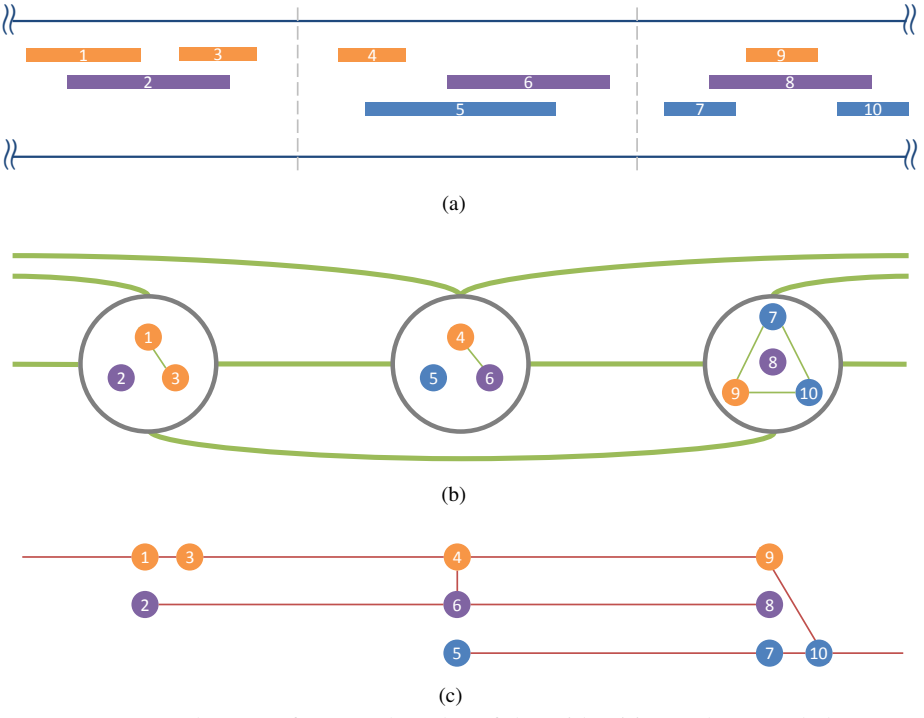

Figure 2: Construct the MST from track nodes of three identities (color-encoded as orange, purple, and blue). (a) Knot construction: Track nodes are organized into three knots (separated by the dotted lines). (b) k-knot graph: The thin green lines represent the edges between track nodes in a knot, and any pair of track nodes from each of the two knots linked by the bold green lines is connected by an edge. (c) MST: Edges of large distances in the k-knot graph are removed. Hence, track nodes of the same identity are more likely to be connected since their associated edges have relatively small distances.

2.2 Construction of Track Nodes

A track node can acquire feature modalities from both face and person tracks, which constitute a stronger representation than individual tracks. Besides, a track node typically has longer presence in timeline than its individual tracks, which allows the uniqueness constraint to be exploited over a longer period. Although a track node provides a unified representation, an erroneous matching of face and person track is irreversible. Besides, any error in the construction of track nodes causes immediate performance degradation in identification since face and person tracks inherit the identity of track node. Thus, we propose the following two-step procedure to construct the track nodes:

1) We use the Hungarian algorithm [1] to match the bounding boxes of face and person tracks in each frame. The Hungarian algorithm takes the cost matrix as input and outputs the matching status between the bounding boxes of face and person track in each frame. Each entry of the cost matrix corresponds to the distance between the center of face bounding box and the hypothesized center of the face according to the location of the person bounding box. If more than half the number of co-occurrence of bounding boxes of a face track and a person track are matched by the Hungarian algorithm, the face and person track are linked.

Figure 3: Face tracks (blue and green) and a person track (red) are merged into one track node.

2) Each face and person track is initially treated as a track node. Two track nodes are merged into a larger track node if there is a linkage between two track nodes. This merging procedure is performed iteratively until it converges. Figure 3 illustrates a track node consisting of face and person tracks.

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. 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.

2.3 Construction of Knots

Structural analysis of video can enhance the performance of identification. For instance, alternating shots [15] are common when filming conversations between two characters. This evidence can be utilized in identification by accumulating the decision for instances appearing in highly-correlated backgrounds. Nevertheless, the video structure usually depends on the media content, and it can be difficult to analyze a long shot. We propose to organize the track nodes into several knots. A knot is defined as the minimum set of track nodes with dependency in a temporal window such that there is no temporal dependency between any two knots. Track nodes can be organized into several knots using the following procedures:
1) We initialize a knot with a track node.
2) We iteratively augment other track nodes that share at least one common frame with any track node in this knot until it converges.
3) Construct another knot by going back to 1) until all the track nodes are organized.
In Figure 2(a), several track nodes are organized into three knots separated by dotted lines.

2.4 Construction of the k-Knot Graph

The k-nearest neighbor (k-NN) graph has been widely used to explore the latent structure of data. However, it does not consider the temporal structure of track nodes and the contextual information of the unique presence of a subject. Hence, we represent the structure of track nodes with an undirected k-knot graph to exploit the contextual information of videos. The k-knot graph is constructed as follows:

1) Any two track nodes within a knot are connected with an edge if both track nodes do not share any common frame. This ensures two track nodes appearing in the same venue are set far apart while exploring the latent structure of track nodes.
2) As two track nodes from each of the two knots do not have temporal dependency, information can be transferred between them. Nevertheless, two track nodes become irrelevant if they are separated by a long temporal duration. In order to emphasize the information of a track node within a short temporal duration, a track node from the i^{th} knot will only be connected with track nodes from the $(i-k)^{th}$, $(i-k+1)^{th}$, ..., and $(i+k)^{th}$ knot. This allows the information to be successfully conveyed among track nodes for identification.

Figure 2(b) illustrates an example for constructing the k-knot graph from the track nodes. The distance between the i^{th} and j^{th} track node in the k-knot graph is defined as

$$d(i,j) = (1-\gamma)d_f(i,j) + \gamma d_p(i,j), \tag{1}$$

where γ controls the tradeoff between the distance induced by face and clothing modality. The $d_f(i,j)$ is the distance between the i^{th} and j^{th} track node induced by the face modality, which is defined as

$$d_f(i,j) = \begin{cases} \min\limits_{m\in F_i, n\in F_j} d_f(\mathbf{x}_i^m, \mathbf{x}_j^n) & \text{, if } F_i \neq \emptyset, F_j \neq \emptyset \\ d_f^{max} & \text{, otherwise,} \end{cases} \tag{2}$$

where F_i is the index set of face features in the i^{th} track node, and $\mathbf{x}_i^m$ represents the m^{th} face feature vectors of the i^{th} track node. If the track node i or j lack the modality from face tracks, $d_f(i,j)$ will be set equal to d_f^{max}, which is the maximum value of $d_f(i,j)$. $d_f(\mathbf{x}_1,\mathbf{x}_2)$ denotes the cosine similarity [12] between $\mathbf{x}_1$ and $\mathbf{x}_2$, and sophisticated metrics, such as the one discussed in [9], can be utilized to improve the performance.

The clothing feature is the RGB color histogram computed from the bounding box corresponding to the torso in the person track. Similarly, we define the distance between the i^{th} and j^{th} track node induced by the clothing modality as

$$d_p(i,j) = \begin{cases} \min\limits_{m\in P_i, n\in P_j} d_p(\mathbf{h}_i^m, \mathbf{h}_j^n) & \text{, if } P_i \neq \emptyset, P_j \neq \emptyset, \\ d_p^{max} & \text{, otherwise,} \end{cases} \tag{3}$$

where P_i is the index set of clothing features in the i^{th} track node. The $\mathbf{h}_i^m$ denotes the m^{th} clothing feature vector of the i^{th} track node. $d_p(\mathbf{h}_1,\mathbf{h}_2)$ represents the chi-squared distance between histogram feature $\mathbf{h}_1$ and $\mathbf{h}_2$. Note that $d_p(i,j)$ will be set equal to the maximum distance d_p^{max} if track node i or j lack the clothing modality.

2.5 Construction of the Minimum-distance Spanning Tree (MST)

The construction of MST automatically removes the unwanted edges of large distance such that the total distance of the spanning tree is minimized. Motivated by this fact, we construct the MST from the k-knot graph to explore the structure of track nodes. In Figure 2(c), we observe that track nodes of the same identity are closer in MST since edges of large distance are removed during the construction of MST. Note that the distance between two track nodes in MST is defined as the summation of distances along the edges connecting these two track nodes. Hence, we define $D(i,j)$ as the distance between the i^{th} and j^{th} track node in MST.

Note that $D(i,j) = d(i,j)$ if the i^{th} and j^{th} track node are directly connected by an edge in MST. We define the similarity between the i^{th} and j^{th} track node as

$$S(i,j) = \exp\left(-\frac{D(i,j)}{\sigma}\right), \tag{4}$$

where σ is the parameter to adjust the similarity.

Let $C_i(y)$ represent the cost for the i^{th} track node if it is treated as identity $y \in \mathcal{C}$, where $\mathcal{C} = \{1,2,\ldots,c\}$ is the identity set. The modeling of $C_i(y)$ will be discussed in Section 2.7. Following the non-local cost aggregation framework presented in [21], the aggregated cost of the i^{th} track node is computed by

$$C_i^A(y) = \sum_j S(i,j)C_j(y) = \sum_j \exp\left(-\frac{D(i,j)}{\sigma}\right) C_j(y). \tag{5}$$

The aggregation procedure can be treated as a filtering operation, where each track node contributes to the task of identification via similarity weighting. Identification becomes robust since information from adjacent track nodes is utilized as a whole for determining the identity. Although the cost aggregation in (5) requires the weighted summation across all the track nodes, Yang [21] provides a linear time exact algorithm to significantly reduce the computational burden.

2.6 Cost Minimization per Knot

The identity of the i^{th} track node can be obtained by assigning the identity that minimizes its aggregated cost in (5). Since a knot consists of several track nodes with temporal dependency, the identity of track nodes from a knot should be jointly determined. The solution can be obtained by enumerating all combinations of labeling that are consistent with the uniqueness constraint such that the aggregated cost of knot is minimized. Hence, we can predict the identities of track nodes in the j^{th} knot by solving

$$\hat{\mathbf{y}}_j = \arg\min_{\mathbf{y}\in\mathcal{Y}_j} \sum_{i\in O_j} C_i^A(y_i), \tag{6}$$

where O_j is the set containing the indices of track nodes in the j^{th} knot. The identities of track nodes in the j^{th} knot form a column vector $\mathbf{y}$, and $\mathcal{Y}_j$ is the set consisting of all the combinations of identities that satisfy the unique presence of an identity. This combinatorial problem can be solved by an optimization procedure based on relaxation technique discussed in [15]. Once the identities of track nodes are determined, the face and person tracks inherit the identity of the track node it is associated with.

2.7 Modeling the Cost Function

The cost function returns the amount of the deviation from the designated subject. Herein, we define the cost for treating the i^{th} track node as identity y as

$$C_i(y) = \begin{cases} -r_i(y), & \text{if } F_i \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Note that the i^{th} track node has cost equal to 0 if it lacks the face modality, i.e. $F_i = \emptyset$. Hence, track nodes that miss face modality will passively receive the information propagated from

adjacent track nodes. Without loss of generality, the unknown class is regarded as the c^{th} class, and $r_i(y)$ in (7) is modeled as

$$r_i(y) = \begin{cases} \frac{1}{|F_i|} \sum_{m \in F_i} \Phi_y(\mathbf{x}_i^m), y \in \{1,2,\ldots,c-1\}, \\ \frac{1}{|F_i|} \sum_{m \in F_i} \lambda \min_{j \neq c} (1 - \Phi_j(\mathbf{x}_i^m)), y = c, \end{cases} \tag{8}$$

where $\Phi_y(\mathbf{x})$ returns the probabilistic output from a support vector machine (SVM) [3, 10]. We follow the setting in [15] to train the SVMs with a second-order polynomial kernel. The training data of the first $(c-1)^{th}$ classes are used to train $c-1$ classifiers using one-versus-all SVM. Note that we do not explicitly model the unknown class since the number of face tracks corresponding to the unknown class is usually insufficient to model the unknown class. Hence, we use the minimum complementary of the probabilistic output among $c-1$ subjects to model the likelihood of unknown class. However, this excessively biases towards the unknown class as the minimum complementary of the probabilistic output can be large for unseen data. We use λ to adjust the likelihood of the unknown class, and λ is obtained from classifying the validation data such that the classification accuracy is maximized. The validation data consists of a subset of training samples from major characters and all the training samples of the unknown class.

3 Experimental Results

3.1 Datasets

We use two datasets provided by the authors of [2, 15] for evaluation.[1] The first dataset consists of 6 episodes of *Big Bang Theory* (BBT), and the second dataset consists of 6 episodes of *Buffy the Vampire Slayer* (BF). We use the face features readily provided in these datasets. The dimension of the feature vector is 240, and the feature coefficients are computed using block-based discrete cosine transform (DCT) from face regions of 48×64 pixels. The BBT dataset provides face and person tracks, while the BF dataset only provides face tracks. Moreover, only 22 % (recall) face tracks are labeled via matching the name of the transcript with the face track that is speaking [2, 7]. These face tracks are weakly labeled with 87 % accuracy (precision) due to the falsely detected speaking face and the mismatch of transcripts. Note that we do not specifically handle the potentially erroneous labeling and use all the available labels for training. More sophisticated methods, reported in [2, 18], can be utilized to further improve the identification performance.

The identification accuracy of face (person) track is computed as the number of correctly identified face (person) tracks over the total number of face (person) tracks in each episode. For comparison, we follow the same setting reported in [2]. There are 11 and 28 subjects in the BBT dataset and BF dataset, respectively. Each dataset has an additional unknown class. Characters that do not belong to any subjects are regarded as belonging to the unknown class, and the uniqueness constraint is not applied to the unknown class. Since the BF dataset does not provide the person track, we compute the clothing features from a hypothesized rectangular region below the face. A similar procedure is also conducted in [6] to extract the clothing features. Hence, the BF dataset demonstrates another scenario where person tracks are not available, and each face track is trivially treated as a track node. In all the experiments, we use $\gamma = 0.8$, $\sigma = 0.1$, and $k = 10$.

[1] Dataset is available at http://cvhci.anthropomatik.kit.edu/projects/mma.

3.2 Discussion on the Experiments

We compare our method with the person identification framework based on MRF [15], which takes the probabilistic output from a trained classifier using semi-supervised learning with constraints (SSLC) [2]. Our trained classifier is denoted as SVM. The performance of identification evaluated on "track node" and "track node with cost aggregation" is denoted as TN and TN+CA, respectively. Considering the uniqueness constraint, we denote TN+CA+K as "TN+CA with cost minimization per knot". Based on the experimental results, we make the following observations:

1. In Table 1, the SVM gives identification accuracy of 78.21% for the BBT dataset. Specifically, the construction of track nodes provides 0.85% improvement. It provides a slight improvement since face tracks belonging to the same track node can be fused for reliable identification. The cost aggregation procedure provides additional 4.49% improvement since the filtering operation of cost aggregation can propagate the information to track nodes without face modality and suppress the impact of noisy instances. Considering the uniqueness constraint, we use (6) to jointly determine the identities of track nodes of each knot. Overall, our method (TN+CA+K) outperforms the MRF framework with SSLC (SSLC+MRF) by 3.48%. The confusion matrix for the identification of face tracks in the BBT dataset is presented in Figure 4, which shows that the proposed method is effective in classifying all the characters including the guest characters. Note that Doug and Summer are not correctly identified since no weakly-labeled face track is associated with these two subjects for training.

2. In Table 2, TN + CA significantly improves the identification accuracy of person tracks over TN, which shows that the identification result of face tracks is successfully transferred to the person track via the cost aggregation on MST. When the uniqueness constraint is considered, TN+CA+K attains the identification accuracy of 86.66%.

3. Since the BF dataset does not provide person tracks, each face track is treated as a track node. Therefore, the performance of SVM is identical to that of TN. In Table 1, the proposed method (TN + CA) achieves the identification accuracy of 69.39%. However, enforcing the uniqueness constraint only gives minor improvement. One of the reasons is that the video structure of BF usually has one or two characters in the scene, which does not provide as much contextual information as in BBT. Our method (TN + CA + K) outperforms SSLC and SVM by 3.71% and 4.26 %, respectively.

4. The performance of person identification depends on the recognizing accuracy of face classifier. In order to fairly compare with [15], we evaluate the identification accuracy of person tracks by providing the groundtruth for face tracks. Following the protocol in [15], we use the BBT dataset to evaluate the character identification. Only the five main characters and the additional unknown class are evaluated. As we can observe in Table 3, our method (TN+CA+K) achieves 4.5% improvement over [15]. This shows that our method performs better than [15] since the information from face and person tracks is utilized as a whole for identification.

5. Table 4 shows the statistics of track nodes in the BBT dataset. In order to investigate the quality of track nodes, we verify whether the face and person tracks in a track node have the same identity using the groundtruth. A track node with any inconsistent identities among its face and person tracks is regarded as an erroneous track node. It is clear that erroneous track nodes only account for a small portion (0.4%) of all the track nodes. In contrast to the MRF framework proposed in [15] where the identities of face tracks are first recognized and transferred to person tracks without visible face based on the affinity of the clothing appearance, our method ensures that all the information is utilized as a whole.

Episode	BBT-1	BBT-2	BBT-3	BBT-4	BBT-5	BBT-6	BBT-Avg.	BF-1	BF-2	BF-3	BF-4	BF-5	BF-6	BF-Avg.
SSLC [0]	89.23	89.20	78.47	76.59	75.09	68.05	79.44	71.99	61.27	66.60	67.07	69.59	61.72	66.37
SSLC + MRF [0,]	95.18	94.16	77.81	79.35	79.93	75.85	83.71	–	–	–	–	–	–	–
SVM	87.94	85.84	77.81	76.25	72.76	68.66	78.21	69.63	62.20	64.20	67.07	69.34	62.45	65.82
TN	90.35	87.26	78.47	77.11	72.04	69.15	79.06	69.63	62.20	64.20	67.07	69.34	62.45	65.82
TN + CA	92.28	91.15	82.22	84.85	78.85	71.95	83.55	74.08	62.31	67.81	72.10	75.95	64.11	69.39
TN + CA + K	94.21	92.39	84.01	87.78	83.15	81.59	87.19	75.65	64.38	66.70	72.81	76.97	63.93	70.08

Table 1: Identification accuracy of face tracks in BBT and BF datasets.

Episode	1	2	3	4	5	6	Avg.
TN	78.54	73.04	75.58	63.77	64.07	61.49	69.42
TN + CA	89.12	87.77	83.67	80.97	78.31	71.01	81.81
TN + CA + K	91.80	89.81	87.71	86.61	82.95	81.09	86.66

Table 2: Identification accuracy of person tracks in the BBT dataset.

Episode	1	2	3	4	5	6	Avg.
MRF [ ]	98.3	89.9	94.8	89.1	85.3	88.5	91.0
TN	86.0	82.6	86.9	78.2	79.1	76.7	81.6
TN + CA	95.7	93.4	96.4	91.5	88.9	85.2	91.8
TN + CA + K	97.8	96.9	97.4	94.2	94.5	92.5	95.5

Table 3: Identification accuracy of person tracks given the groundtruth identities of face tracks in the BBT dataset.

Although body-only track nodes lack face modality, they serve as the relay for propagating the inference of other track nodes. Moreover, the duration of the track node accounts for the temporal appearance of an identity in the timeline, and thus the pairwise constraint between track nodes is generally stronger than just the face or person track alone.

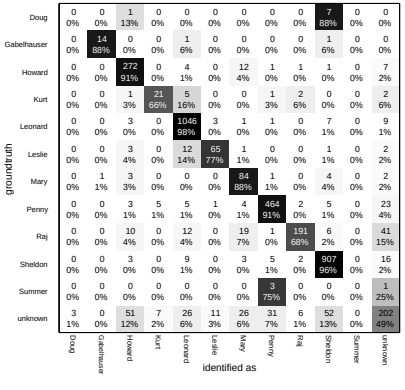

Figure 4: Confusion matrix over Episode 1-6 of BBT for TN + CA + K.

Episode	1	2	3	4	5	6
# Face tracks	622	565	613	581	558	820
# Person tracks	671	638	643	657	604	883
Our track nodes (TNs)						
# Face-body TNs	527	469	476	481	424	604
# Face-only TNs	14	21	66	43	80	156
# Body-only TNs	142	168	158	174	174	262
# All TNs	683	658	700	698	678	1022
# Erroneous TNs	0	1	1	2	6	11
# Knots	362	348	375	330	281	316

Table 4: Statistics of the tracks, track nodes, and knots in Episode 1-6 of BBT.

4 Conclusions

We propose a unified framework for character identification in a TV-series. We construct the track nodes from face and person tracks, and the track nodes serve as the basic unit in constructing the MST. Hence, track nodes with similar appearance are adjacent in MST. Then non-local cost aggregation is performed on the MST, which serves as a filtering operation to suppress the impact of noisy instances and provides the inference to track node without face modality. Considering the unique presence of a subject, the identities of track nodes with temporal dependency can be jointly determined by minimizing the aggregated cost of those track nodes. Experimental results confirm the effectiveness of our method.

5 Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon.

References

[1] R. Ahuja, T. Magnanti, and J. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[2] M. Bäuml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[5] 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.

[6] M. Du and R. Chellappa. Face association across unconstrained video frames using conditional random fields. In *European Conference on Computer Vision (ECCV)*, 2012.

[7] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy - Automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006.

[8] 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.

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

[10] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, Oct. 2007.

[11] 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.

[12] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[13] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *European Conference on Computer Vision (ECCV)*, 2014.

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

[15] 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.

[16] 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.

[17] 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.

[18] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Improved weak labels using contextual cues for person identification in videos. In *IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG)*, 2015.

[19] 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.

[20] 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.

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