"Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series

Makarand Tapaswi Martin Bäuml Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{makarand.tapaswi, baeuml, rainer.stiefelhagen}@kit.edu

Abstract

We describe a probabilistic method for identifying characters in TV series or movies. We aim at labeling every character appearance, and not only those where a face can be detected. Consequently, our basic unit of appearance is a person track (as opposed to a face track). We model each TV series episode as a Markov Random Field, integrating face recognition, clothing appearance, speaker recognition and contextual constraints in a probabilistic manner. The identification task is then formulated as an energy minimization problem. In order to identify tracks without faces, we learn clothing models by adapting available face recognition results. Within a scene, as indicated by prior analysis of the temporal structure of the TV series, clothing features are combined by agglomerative clustering. We evaluate our approach on the first 6 episodes of The Big Bang Theory *and achieve an absolute improvement of 20% for person identification and 12% for face recognition.*

1. Introduction

This paper addresses the problem of automatic labeling of all characters in a TV series. Person identification has direct applications in the generation of meta-data for use in indexing and fine-grained retrieval of specific scenes ("Show me shots with Sheldon"). More importantly, person identification forms the basis for other types of multimedia analysis that benefit from person-specific models.

Albeit the number of main characters in a TV series is typically low (usually below 15), the recognition problem is challenging due to high variability in view points, facial expressions, general appearance and lighting conditions, as well as occlusions, rapid shot changes and moving cameras. State-of-the art approaches [7, 14] tackle the problem by mainly relying on face tracks. However, we think it is desirable to identify the characters also when their face is not visible. Our goal in this paper is to develop a method to identify *all* character appearances in a TV series or movie, specially where the face cannot be detected or tracked. Accordingly, we aim to assign identity labels to all *person* tracks, instead of face tracks. This introduces an additional challenge that we now need to infer an identity even when the face is not observable. Furthermore, we wish to integrate different features in a principled way, and restrict not only to facial and clothing appearance, but also use constraints, such as "the same person cannot appear twice in the same frame". Towards this goal we model the character appearances as a Markov Random Field (MRF) and integrate cues from face, speech and clothing in a common framework. In addition, we leverage structural elements that are common to TV series: (i) the division of the plot in *scenes* which we use as hints to re-learn clothing models and (ii) the concept of *alternating shots*, where the sequence alternates between two shots, *e.g.* in a conversation, which effectively allows us to use additional evidence for the identity decision.

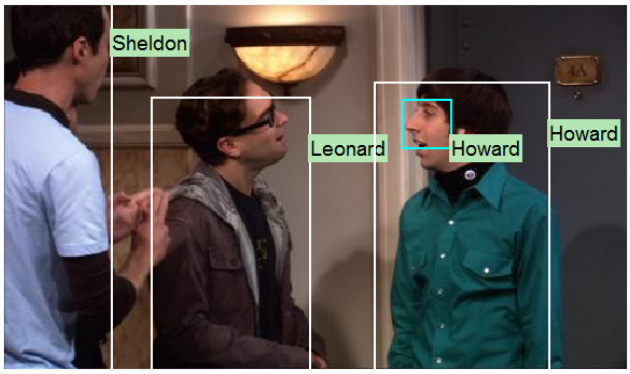

Figure 1: A sample image from *The Big Bang Theory*. Note how our approach is able to correctly identify all three characters by relying on clothing when the face is unseen.

We demonstrate our approach on episodes 1-6 of season 1 of the sitcom *The Big Bang Theory*. Each episode is about 20 minutes in length. Figure 1 shows a sample image.

1.1. Related work

Previous work on identifying characters in movies and TV series can be roughly divided into two categories: (i) person retrieval, where the goal is to find all occurrences of a character from a given example image or sequence and (ii) full labeling of every character appearance with a unique identity. Our work falls into the latter, therefore we focus on

a discussion of relevant work in this area.

Closest to our work are the works of Everingham *et al*. [7], Sivic *et al*. [14] and Ramanan *et al*. [12] which have the goal of labeling every character appearance in videos. Everingham *et al*. [7] propose a method for unsupervised labeling of frontal face tracks. They build exemplar sets for all characters using subtitles in conjunction with aligned transcripts. All remaining face tracks are matched against these exemplar sets in a nearest neighbor fashion. Clothing features are computed from a box beneath the face, to support labeling decisions when the face descriptors are indecisive, *e.g*. due to differences in lighting and pose. Sivic *et al*. [14] extend [7] to half- and full-profile faces, thus increasing the coverage of the labeled persons. Köstinger *et al*. [10] take advantage of semi-supervised multiple instance learning to incorporate weakly labeled faces during training and show improved performance on the same data set. In contrast to [7], Cour *et al*. [4] propose a method for labeling characters without transcripts and use in-video dialog cues to capture references to identities. Ramanan *et al*. [12] cluster faces in a hierarchical procedure using clothing and hairstyles as additional cues. Some of the obtained face clusters are labeled manually, and are used to identify all face tracks using a nearest neighbor fashion similar to [7].

However, [7, 10, 12, 14] all have one common drawback. They are limited to character occurrences for which a face is detectable. Further, since the identification is performed for individual tracks, constraints such as "the same person cannot appear twice in one frame" cannot be integrated.

MRFs have been successfully applied to person identification in photo albums. Anguelov *et al*. [1] perform recognition primarily based on faces, and incorporate clothing features from a region below the face. Photo albums are divided into events, during which people are expected to have similar clothing. They also model the constraint that two people appearing in the same photo cannot be assigned the same identity. However, this approach is also limited to occurrences where a face can be detected.

1.2. Overview of our approach

In our approach, we build on ideas from the discussed related work and extend them to labeling of full persons in videos in a probabilistic framework.

Our approach can be divided into three steps. We first start with basic video analysis, where we split the video into scenes and shots, and detect alternating shots (Sec. 2). In the second step (Sec. 3), different cues give us strong and weak hints on the character identities: We first track both faces and full persons in each shot. We then perform face recognition on the face tracks, match clothing for the person tracks and also perform speaker identification. Finally, the outputs of the individual components and additional constraints are integrated in an MRF, and the labels for person tracks are obtained by energy minimization (Sec. 4). We present experimental results in Sec. 5 and conclude with a discussion of the approach and possible future work in Sec. 6.

The main contributions of our work are: (i) We achieve full coverage by labeling person tracks instead of face tracks only. To deal with the problem of labeling tracks without faces, we propose a method for unsupervised learning of clothing models. (ii) We detect and leverage structural elements that are commonly found in TV series, namely scenes and alternating shots, for learning the clothing models and incorporating additional constraints, respectively. (iii) The speech modality is included in our model without the need for subtitles or fan transcriptions. The ambiguity of assigning a track to a recognized speaker is handled by the introduction of a latent presence variable.

2. Structural video analysis

We simplify the problem of labeling the full video by first splitting it into scenes and shots, which can, to some extent, be treated independently. Additionally, we obtain important information about the structure of the video which is leveraged for learning clothing models and adding supplementary constraints to our model.

2.1. Shot boundaries

A normalized version of the *Displaced Frame Difference* (DFD) [16], the difference between consecutive motion compensated frames is used to detect the shot boundaries. $DFD(t) = \|F(r,t) - F(r + D(r), t-1)\|$, where $D(r)$ is the optical flow between frames $F(r, t-1)$ and $F(r,t)$. We achieve 1981 correct detections, 2 misses and only 8 false positives over the 6 episodes. The high precision of the shot boundary detector helps our face tracker initiate and terminate tracks. Further, it is also the first step towards reliable detection of alternating shots (Sec. 2.3) and creation of the model (Sec. 4).

2.2. Special sequences

Special transition effects (see Figure 2) or special audio jingles are commonly added to TV series (Seinfeld, Friends, etc.) to indicate a larger change in time and/or location of the plot. We call the set of shots between two such indicators a *scene*. We observe that the clothing of the characters usually remains unchanged *within* scenes, but may change from one scene to another. Thus a scene change is a good juncture for re-learning clothing models.

In *The Big Bang Theory* the background of the special sequence is a colour gradient, which is hard to represent using a few colours. Thus, we are able to detect these sequences reliably by thresholding the difference between each frame and its eight dominant colour representation [11] (19 detections, 0 false positives, 0 misses). For other TV series, depending on whether a special sequence or a special jingle

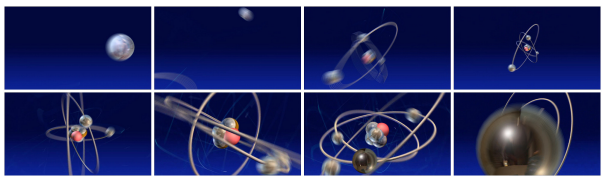

Figure 2: Typical special sequence denoting a scene change.

denotes a scene change, simple video and/or audio template matching should provide sufficient recognition rates.

2.3. Alternating shots

We observe another interesting feature in the temporal structure of the TV series. During conversations, it is seen that shots flip back and forth between actors. Usually the actors seen across these alternating shots remain the same and do not move. This can be included in our model as a constraint on the possible identities within a shot or used to accumulate scores for person tracks across shots.

Alternating shots are detected by applying a threshold on the normalized DFD between the first frame of each shot and its corresponding second-consecutive shot. With this technique, we achieve an equal error rate of 4.5%.

3. Face, clothing and speaker identification

We obtain face and person tracks in each shot. From each track we extract features which are used for the recognition. Speaker identification is also performed for each shot.

3.1. Face detection and tracking

We employ a detector-based multi-pose face tracker [2], incorporated in a particle-filter framework. The tracking is performed in an online fashion, *i.e*. the tracker does not know the detections of the entire shot in advance, but only uses the state of the previous frame to infer the location and head pose of the faces in the current frame. We initialize tracks by scanning the whole image every fifth frame, using frontal, half-profile and profile face detectors [8]. This allows us to detect and subsequently track faces independent of their initial pose. To score a particular particle of a track, we employ a total of 11 face detectors, one for each of the yaw-angles $-90, -60, -45, \ldots, 0(\text{frontal}), \ldots, 45, 60, 90$. The face detectors already achieve a low false positive rate, which is further reduced by subsequent tracking. The tracker runs close to real-time ($\sim$10fps) and on average produces 650 tracks in a typical 20-minute episode.

3.2. Face recognition

We extract features for face recognition from each track (see Figure 3), building on a local appearance-based approach [6]. We first locate the eyes within the tracked face region using an eye detector based on the same features as our face detector. While this is certainly not as accurate as

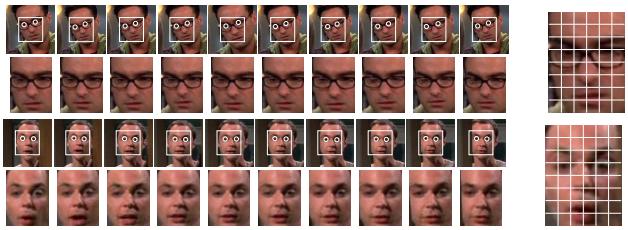

Figure 3: Face tracks, eye detections and aligned and cropped faces for each frame (left). DCT feature computation grid of 6×8 blocks on aligned images (right).

the facial features localization in [7], the detected eye positions suffice to normalize the face to a canonical pose and a standard size of 48×64 pixels. The normalized face is split into 6×8 blocks of 64 pixels each, and the Discrete Cosine Transform (DCT) is computed over each of these blocks. We store the first five coefficients (ignoring the DC value) of each block and obtain $\mathbf{x}_t^{(i)}$, a $6 \times 8 \times 5 = 240$ dimensional feature vector for frame t in track i. For recognition, second order polynomial kernel SVMs are trained in a 1-vs-all fashion for all primary characters j. Normalized classification scores are accumulated over all the frames $t \in T_i$ for each classifier C_j

$$f_j^{(i)} = \frac{1}{|T_i|} \sum_{t \in T_i} \Phi_j(\mathbf{x}_t^{(i)}); \Phi_j(\mathbf{x}_t^{(i)}) = \frac{1}{1 + e^{-\theta_{1j} - \theta_{2j} C_j(\mathbf{x}_t^{(i)})}} . \tag{1}$$

where parameters θ are learned while training the classifiers. The face id score $f_j^{(i)}$ ranges from 0 to 1 and can be interpreted as a pseudo-probability. A score towards the *Unknown* identity is obtained from each frame as the smallest remaining uncertainty considering all known identities:

$$f_U^{(i)} = \frac{1}{|T_i|} \sum_{t \in T_i} \min_j (1 - \Phi_j(\mathbf{x}_t^{(i)})) . \tag{2}$$

We do *not* make a hard decision at this point, but keep all scores and include them as evidence in our model. Thus, even if a face score is not the highest for a given track, it still influences the final identity decision during fusion.

3.3. Person detection and tracking

Person detection is a much harder task than face detection due to the non-rigid nature of the human body and the wide range of general person appearance. In a TV series, both full-body and upper-body shots are common. To ensure both types are detected, we adapt the part-based poselets [3], and use them to track full persons in a detector-based tracker similar to the face tracking approach described in Sec. 3.1. Owing to the difficulty of the problem, our tracker is able to obtain about 47% recall, with a precision around 77%. In order to assess our approach independent of the quality of the person tracker, the undetected person tracks are manually annotated.

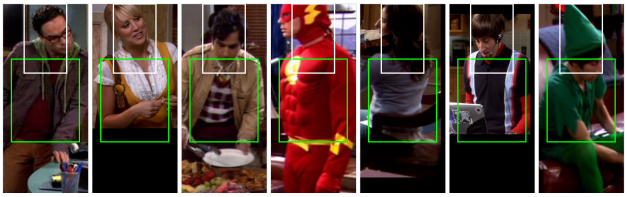

Figure 4: Sample poselet-based person detections (1-5), manual annotations (6-7). The green box is the clothing feature area, the white denotes the region where we search for a face detection.

3.4. Clothing clustering and identification

Clothing can be an important cue to disambiguate different people and has been used successfully to support face recognition (see, *e.g.*, [1, 7, 14]).

We propose a novel method to learn character-specific clothing models in an unsupervised fashion using the output of the face identification from Sec. 3.2. Our features are RGB histograms computed from a rectangular region located within the *person* bounding box (therefore being independent of face detection). We show that this simple segmentation and feature work well. Of course it is possible to segment the person more accurately (as in [9]) and/or use more sophisticated features instead, and our method will directly benefit from it. Figure 4 shows person detections and the corresponding clothing boxes (in green) and search areas for face detections (in white).

We need to tackle some problems when learning clothing models: (i) Characters change clothing within an episode, so we need to detect when to re-learn clothing models. (ii) Clothing features are unreliable, *i.e.* the inter-person difference can be smaller than intra-person variance due to similar clothing of different characters, illumination or pose changes. (iii) Face id decisions are unreliable, *i.e.* we can only expect that around 70-80% of the face identities are correct, and (iv) not every person detection/track has associated face information.

We solve the first problem by learning clothing models for each scene as demarcated by the special sequences from Sec. 2.2. While in some cases we re-learn the clothing model unnecessarily, it is not harmful either. In practice, every character to be recognized appears in the scene at least once with their face visible. This allows us to learn a new character-specific clothing model for each scene. The learning algorithm (see Figure 5) consists of three steps:

1. Clustering: We cluster all clothing features within one scene agglomeratively using ward-linkage [15] over the Euclidean distance. By choosing a low cut-off threshold θ_c (*e.g.* at 6% of the maximum distance), the clustering yields many more clusters than the number of characters in the scene. In this way the clusters remain relatively homogeneous, and it is unlikely that features from two different characters mix (thus dealing with issue (ii)). Note that we still do not know to whom the clusters belong.

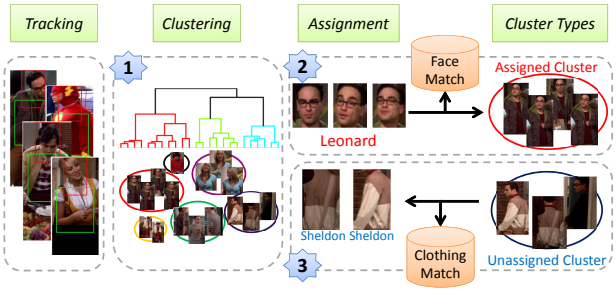

Figure 5: Overview of clothing model learning and recognition.

2. Transfer face results: For some of these clustered person detections we have faces found by our face tracker (lying within the white box in Figure 4). If a face is found, we associate the person's identity provided by the face recognizer with the feature within the cluster. We then categorize clusters into two types. The clusters which have person detections with sufficient number of faces, of which at least θ_{id} (60%) are recognized as the same character are called *assigned clusters*. This addresses problem (iii) that not all faces are correctly classified. An assigned cluster is tagged with the identity of the most common face result, and each character is assigned scores according to the distribution of face results within the cluster.

3. Clothing classification: Finally we tackle problem (iv) – the case of person detections without faces – by scoring features in unassigned clusters individually. Each such clothing feature is matched to the nearest means of the assigned clusters. Consider an assigned cluster k, with identity $\gamma(k)$, having N_k features. Its mean feature is $\overline{h}_k = \frac{1}{N_k}\sum_j h_{jk}$, where h_{jk} is the histogram of detection j. Then for feature l in an unassigned cluster m, the identity assignment is done according to

$$\gamma_l(m) = \gamma(k^*),\ k^* = \arg\min_k \|h_{lm} - \overline{h}_k\| \,. \qquad (3)$$

Thus, for features in the assigned clusters, a score vector is the distribution of face results within the cluster, while for features in unassigned clusters, the distances to each of the assigned clusters in that scene are used to generate a score. Finally, the scores of individual features in one person track are accumulated to generate a score for the person track as a whole. Owing to the soft assignment of scores, we still retain the possibility to overturn the decision when integrating the clothing result in the global model.

3.5. Speaker identification

Our speaker identification system is based on [13]. We build a Gaussian Mixture Model (GMM) for each primary character, and one to encompass all other non-primary speakers. The audio is first down-sampled and split into overlapping frames of 20ms. For each frame, Mel Frequency Cepstral Coefficients [5] are computed and are used

to train the GMMs using Expectation Maximization. For identification, we use maximum a-posteriori probability

$$\hat{S}(j) = \arg\max_{1 \le i \le N} p(\lambda_i | X(j)) \; . \tag{4}$$

where $X(j)$ is the feature for frame j, λ_i is the model for the ith speaker, and $\hat{S}(j)$ the estimated speaker. Spurious errors are removed by mode-filtering. Finally, we say that a character is speaking within a shot if his largest speech segment is longer than 25% of the shot duration.

It is not easy to associate the speaker with one of the appearing characters in the shot. [7] and [14] try to determine the current speaking character from lip movements. We observe that the speaker usually appears in the shot and therefore impose the constraint that one of the appearing characters should match each identified speaker.

4. Global model

We model our problem of automatic labeling of characters in video as a Markov Random Field. This effectively combines the individual face, clothing and speaker modalities and also allows to easily include contextual constraints.

Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote the set of shots within a scene. For shot S_i, we have a set $\mathcal{M}_i = \{\mu_{i1}, \ldots, \mu_{in}\}$ of identity variables, one for each of the n person tracks in the shot. $\mathcal{P}^{(i)}$ denotes a latent variable for the presence of characters in shot S_i. We associate the clothing results c_{ij} and face results f_{ij} (only if present) with identity variables μ_{ij} and the speaker recognition results s_i with the presence variable $\mathcal{P}^{(i)}$. For our choice of TV series, f_{ij}, c_{ij}, s_i and μ_{ij} are 6-dimensional vectors comprising the recognition scores for five main characters and the *Unknown* category.

For alternating shots, we first associate tracks across shots by checking for large overlap in tracked area. Typically, since motion between alternating shots is minimal, we can assume that the characters stay in their same place. The clothing recognition scores for tracks which match are accumulated and normalized. If $\{\ldots, i-2, i, i+2, \ldots\}$ is a sequence alternating shots, then the new clothing score for track j is $c_{ij} = \ldots + c_{i-2,j} + c_{i,j} + c_{i+2,j} + \ldots$.

The MRF defines a joint probability over the identity and presence variables. For the task of identity labeling, we are interested in the maximum a-posteriori assignment of the identity variables given the face, clothing and speaker results. To describe the relationships between the random variables, MRFs can be approached from two perspectives – a probability maximization problem or an energy minimization problem. The latter is chosen in this paper.

4.1. Energy functions

From intuition, we want the energy of the system to reduce when the computed identity of the person matches that of the face and clothing. We would like to increase the energy if the identified speakers are not among the labeled identities. Furthermore, when two simultaneously appearing tracks are assigned the same identity, it should also increase the energy of the system. This gives us 4 energy-terms: a *clothing energy* E_C, a *face energy* E_F, a *speaker energy* E_S and a *uniqueness energy* E_U. Figure 6 depicts the graphical model along with the energy terms.

We define the clothing energy between the identity variable μ_{ij} and its associated clothing result c_{ij} as

$$E_C(i,j) = -\langle \mu_{ij}, c_{ij} \rangle \; . \tag{5}$$

The inner product matches our initial intuition that the energy decreases when the identity assignment agrees with the clothing result. Similarly, we compute the face energy as

$$E_F(i,j) = -\langle \mu_{ij}, f_{ij} \rangle \; . \tag{6}$$

Further, for each person track, we introduce a *regularization energy* denoted by

$$E_R(i,j) = \langle \mu_{ij}, \mu_{ij} \rangle \; . \tag{7}$$

which prevents the μ_{ij} from growing too large. We weight the three energies by w_C, w_F and w_R respectively. This modality energy for shot S_i over all person tracks j is

$$E_{FCR}(i) = \sum_j w_F E_F(i,j) + w_C E_C(i,j) + w_R E_R(i,j) \; . \tag{8}$$

From the identity variables, we deduce the presence of characters appearing in one shot $\mathcal{P}^{(i)} = \phi(\sum_j \mu_{ij})$, where $\phi(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. The sigmoid ensures that the presence term is not biased towards one character when there are multiple tracks of the same actor in the same shot (*e.g.* due to occlusions or failures in the tracker). Given the presence term $\mathcal{P}^{(i)}$ for a shot S_i, we capture the speaker penalty as

$$E_S(i) = \langle (1 - \mathcal{P}^{(i)}), s_i \rangle \; . \tag{9}$$

which says that if the speaker is not among the people present in the shot, then induce a penalty.

Finally, we add the uniqueness penalty as a pairwise energy between combinations of identity variables in one shot. Let $\mathcal{T}_i$ be the set of person track pairs which share at least one common frame in shot S_i. Note that two or more *Unknown* characters may appear together, and should not be penalized. Hence we consider $\mu'_{ij} = \mu_{ij,1:D}$, where D is the number of main characters for the uniqueness penalty. The uniqueness energy for shot S_i is therefore defined as

$$E_U(i) = \sum_{(j,k) \in \mathcal{T}_i} \langle \mu'_{ij}, \mu'_{ik} \rangle \; . \tag{10}$$

Combining the modality and constraint energies, we obtain

$$E(i) = E_{FCR}(i) + w_S E_S(i) + w_U E_U(i) \; . \tag{11}$$

and the optimal identity labels are computed by minimizing $E(i)$ jointly over all identity variables in $\mathcal{M}_i$.

$$\mu^* = \arg\min_{\mu} (E(i)) \; . \tag{12}$$

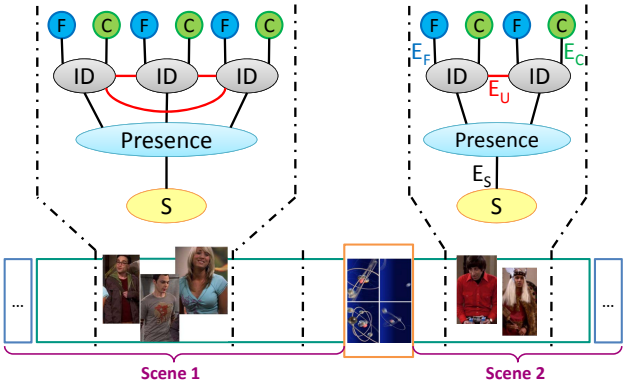

Figure 6: Graphical representation of the MRF illustrating the division into scenes and shots, and interaction between identity variables μ_{ij} with face and clothing modalities. The speaker information is included via the presence node $\mathcal{P}^{(i)}$, and uniqueness is enforced by potentials (red links) between the identity variables.

For our experiments, we use equal weights for the face, clothing and regularization energy $w_C = w_F = w_R = 1$, while speaker penalty and uniqueness are enforced more stringently with $w_S = w_U = 2$.

5. Experimental results

We evaluate our approach on the first six episodes of the first season of the popular American sitcom *The Big Bang Theory*[1]. Table 1 provides some statistics about the data set. Note that of all person tracks (#Tracks), about 30% do not contain faces (#with Faces) and are missed if face detection is used as the sole person detection scheme. We also show the number of tracks which belong to the main characters (#Main), which are the focus of our identification scheme. All other tracks are categorized as *Unknown*. The amount of speech (in seconds) is also presented.

5.1. Person labeling

We present the results of our experiments in multiple stages of improvement (see Table 2). The face and speaker models are trained on episodes 4, 5 and 6. From these episodes, we count which actor appears most frequently. In the worst-case scenario of absence of any recognition scheme, labeling all tracks as the most likely prior is the best option (here: Sheldon, with 29.7%). This is referred to as the *MaxPrior* baseline, and it captures the difficulty of the data set.

As a second baseline, we use the face id results when a person track can be associated with a face track. When the face is unseen, the track is assigned the label of the most likely character, *Sheldon*. On average, this allows us to correctly label 63.1% of the tracks.

[1]Video events, tracks and identity labels for the data set available at http://cvhci.anthropomatik.kit.edu/projects/mma

Episode	E1	E2	E3	E4	E5	E6	Total
#Tracks	662	616	619	632	573	802	**3904**
#with Faces	499	408	438	444	384	534	**2707**
#Main	636	614	520	455	448	670	**3343**
#Unknowns	26	2	99	177	125	132	**561**
Speech (s)	766	650	699	652	570	621	**3958**

Table 1: Statistics for episodes 1–6 of *The Big Bang Theory*

Episode	E1	E2	E3	E4	E5	E6	Avg
MaxPrior	32.1	26.4	17.2	44.0	23.2	22.1	**27.5**
Face	70.1	64.9	59.6	65.7	57.1	61.3	**63.1**
Clothing	89.7	76.1	78.7	73.1	62.7	77.2	**76.2**
F+C	90.6	79.5	80.5	79.6	68.6	80.3	**79.8**
F+C+S	90.6	80.5	80.5	80.1	68.6	80.3	**80.1**
FullModel	92.5	83.1	80.8	83.4	69.7	85.8	**82.6**

Table 2: *Person* identification accuracy from baseline to fusion

Episode	E1	E2	E3	E4	E5	E6	Avg
Face	81.9	74.4	72.5	76.4	74.3	71.1	**75.1**
Clothing	93.5	81.6	91.1	78.0	76.1	79.8	**83.4**
FullModel	98.3	89.9	94.8	89.1	85.3	88.5	**91.0**

Table 3: *Person* identification acc. with groundtruth face labels

Using clothing instead of faces increases the recognition accuracy to around 76%. This is mainly due to the fact that using clothing it is possible to also identify tracks with an unseen face. Note however that the clothing models are learned using intermediate face results (for assigning cluster labels), so this result depends on the accuracy of the face recognition. Combining face and clothing provides a 3% increase. This shows that joint recognition is fruitful even when clothing models are learned from face results.

Speaker identification is unable to provide a significant improvement (0.3%). One possible reason might be that the speaker constraint is limited to preventing a track from being labeled as a non-present person. It is not capable however to reduce confusion between two persons present in the same shot. By associating the current speaker with one of the person tracks, *e.g.* using speaker detection as in [7, 14], we expect a more significant impact in the future.

Finally, the full model including the uniqueness constraint and alternating shots achieves an average recognition accuracy of 82.6% over all episodes, a further 2.5% improvement over face, clothing and speech combined. The results for all episodes individually and on average are displayed in Table 2.

An interesting aspect is to analyze the influence of the face recognition performance on the subsequent stages. In Table 3 we present results for face, clothing and full model usage on all person tracks when using *groundtruth* for the face labels. Note that the face results are not 100% (but 75.1%) because not all person tracks have associated face tracks (in this case the max-prior *Sheldon* is used). We

Episode	E1	E2	E3	E4	E5	E6	Avg
Face only results, Person tracks							
FAR	69.2	100	85.8	57.6	90.4	74.2	**79.5**
FRR	7.7	3.2	2.1	1.6	0.4	1.4	**2.7**
FCR	21.1	32.5	25.1	17.0	22.9	25.4	**24.0**
Full model results, Person tracks							
FAR	38.5	50.0	60.6	24.3	68.0	15.2	**42.7**
FRR	5.2	6.6	6.5	9.9	15.1	5.9	**8.2**
FCR	1.3	10.6	5.1	4.4	8.4	3.8	**5.6**

Table 4: False Acceptance, Rejection and Classification rates for each episode before and after usage of the model

Episode	E1	E2	E3	E4	E5	E6	Avg
Face	77.5	74.5	68.8	76.9	63.1	69.8	**71.8**
F+C	89.3	84.7	76.4	82.1	67.4	78.8	**79.8**
FullModel	89.3	84.9	80.8	86.7	73.3	84.1	**83.2**

Table 5: *Face* recognition (Sec 5.2) acc. from baseline to fusion

observe that using clothing clustering increases the performance to 83.4%, and the full model further to 91.0%. This shows that better face recognition will consistently improve all stages. However, clothing-based recognition and the full model are essential for labeling all non-face person tracks.

In Table 4 we report the False Acceptance Rate (FAR), False Rejection Rate (FRR) and False Classification Rate (FCR) on the six episodes. Note the large reduction in FCR from 24% to 5.6% by using our proposed scheme. Reduction in FCR is crucial for using the person identification for higher-level semantic tasks. The poor classification of *Unknown* characters can be attributed to lack of explicitly modeling all *Unknown* characters individually. However, *Unknown* characters account for only 14% of all person tracks.

Figure 7 presents precision-recall curves for our test episodes 1, 2 and 3 on person tracks. Finally, in Figure 8 we see sample images from our database where correct person identification is achieved either due to fusion of face, clothing and speech modalities. We also see the uniqueness constraint help resolve confusion.

5.2. Face labeling

In another experiment, we evaluate our approach for the task of the *face* recognition, *i.e*. we now use the model for labeling *face* tracks. Table 5 shows that the addition of clothing information improves performance by 8% and finally fusing speaker information and the uniqueness penalty provides another 4% increase, improving the overall recognition rate from 71.8% to 83.2%.

6. Conclusion

In this paper, we propose to shift the focus in automatic labeling of characters from face recognition to full person recognition. Person tracks are used as the basic unit to label all character occurrences in a TV series. We model the problem as a Markov Random Field, fusing different modalities, face, clothing and speech, efficiently in a probabilistic framework. We analyze the scene structure of the TV series in order to learn appropriate clothing models for reliably identifying person tracks without faces. The MRF also facilitates the use of contextual cues, and can be further extended with other cues, such as gender or hair color, if available. In the future, we also plan to incorporate transcripts and subtitles similar to [7, 14, 10] to allow for fully unsupervised labeling. Although this is out of the scope of the current work, a better person tracker would be desirable.

Acknowledgments We thank Arne Schumann for providing the person tracks. This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and was partially funded by the German Federal Ministry of Education and Research (BMBF) under contract no. 01ISO9052E. The views expressed herein are the authors' responsibility and do not necessarily reflect those of OSEO or BMBF.

References

[1] D. Anguelov, K.-C. Lee, S. B. Gökturk, and B. Sumengen. Contextual Identity Recognition in Personal Photo Albums. In *CVPR*, 2007.

[2] M. Bäuml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *Advanced Video and Signal-Based Surveillance*, 2010.

[3] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

[4] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking Pictures: Temporal Grouping and Dialog-Supervised Person Recognition. In *CVPR*, 2010.

[5] S. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[6] H. K. Ekenel and R. Stiefelhagen. Analysis of Local Appearance Based Face Recognition: Effects of Feature Selection and Feature Normalization. In *CVPR Biometrics Workshop*, 2006.

[7] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is..." Buffy – Automatic Naming of Characters in TV Video. In *BMVC*, 2006.

[8] B. Fröba and A. Ernst. Face Detection with the Modified Census Transform. In *FG*, 2004.

[9] A. C. Gallagher and T. Chen. Clothing Cosegmentation for Recognizing People. In *CVPR*, 2008.

[10] M. Köstinger, P. Wohlhart, P. Roth, and H. Bischof. Learning to Recognize Faces from Videos and Weakly Related Information Cues. In *Advanced Video and Signal-Based Surveillance*, 2011.

[11] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

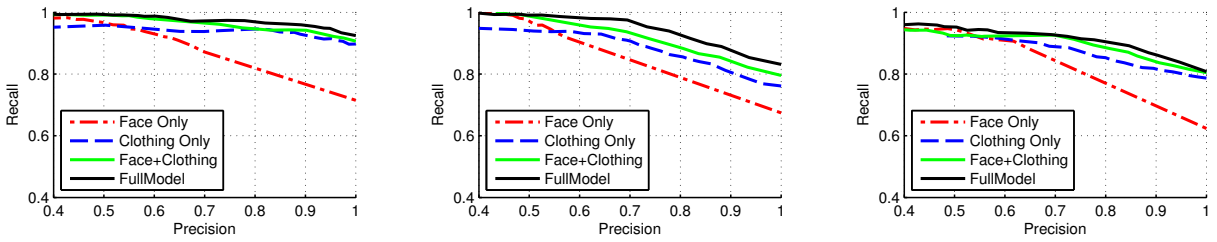

Figure 7: Precision-Recall curves for episodes 1–3 of *The Big Bang Theory*. Recall is the proportion of tracks selected at given confidence threshold, while precision is the ratio of correct tracks to these selected tracks.

Figure 8: Examples of correct character labeling when the face is not visible (row 1), when the face classification is erroneous (row 2), when the clothing classification is erroneous (row 3) and when the result is corrected due speaker penalty (row 4, column 1) or the uniqueness constraint (row 4, columns 2-3). The labels adjoining the detection boxes are as follows: first row denotes the ground truth (yellow background), the second row is the direct face or clothing result and third row our model fusion result. The results have red background when face/clothing/model results are incorrect and green background when correct (best viewed in color).

[12] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *ICCV*, 2007.

[13] D. Reynolds and R. Rose. Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models. *Speech and Audio Processing*, 3(1):72–83, 1995.

[14] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – Learning person specific classifiers from video. In *CVPR*, 2009.

[15] J. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of American Statistical Association*, 58(301):236–244, 1963.

[16] Y. Yusoff, W. Christmas, and J. Kittler. A Study on Automatic Shot Change Detection. *Multimedia Applications and Services*, 1998.