Multi-Person Re-Identification Based on Face, Pose and Texture Analysis in Unconstrained Videos

Jaime Gallego
Eventlab
Universidad de Barcelona
Barcelona, Spain
jgallego@ub.edu

Mel Slater
Eventlab
Universidad de Barcelona
Barcelona, Spain
melslater@ub.edu

Abstract—We present a method for re-identification of images of multiple people that appear in 2D RGB video sequences. The method needs no initialization or supervision and works with unconstrained sequences that include camera shot transitions and strong visual variations. In order to preserve tracking along the frames, our method combines facial recognition, clothing texture analysis and pose detection to compute a distance between tracked people at frame t-1, and detected people at frame t. We use techniques based on Convolutional Neural Networks (CNNs) to extract this information from the people that appear in the images. The results obtained show that the proposed method achieves accurate tracking of the people even in those difficult sequences where faces appear occluded or people present similar textures.

Index Terms—Multi-person re-identification, people tracking, Unconstrained tracking in videos

I. INTRODUCTION

Multi-person re-identification consists in detecting and locating all people that appear in a video sequence, and establishing the correspondences between the people and their images along the frames to maintain their identities. It is a challenging problem in computer vision as one person often appears to look drastically different in multiple shots due to significant variations in scale, pose, expression and illumination.

Person re-identification is a central process for high impact applications such as video understanding and analysis, scene recognition, event detection and human poses extraction. In order to support correct functionality for any kind of video sequences such as tv shows or hand held cameras, we need to deal with unconstrained videos, which have the following main characteristics:

- Presence of camera shot transitions. The contents of two neighboring shots may be completely different.
- Dramatic character appearance variations along the frames due to changes in scale, pose, expression and illumination.

This work is funded by the European Research Council (ERC) Advanced Grant Moments in Time in Immersive Virtual Environments (MoTIVE) number 742989.

II. RELATED WORK

Multi-person re-identification is a major topic that has been extensively studied in the literature. Traditional approaches have been based on hand-crafted features to perform the correspondence between detected people along the frames. Dalal et al. [1] used histograms of oriented gradients (HOG) to construct an appearance model for each target. Some approaches often address multi-face tracking techniques. Zhao et al. [2], presented the Motion Structure Tracker to solve the problem of tracking in very crowded structured scenes. They combined visual tracking, motion pattern learning and multi-target tracking. In [3], the authors presented a Network Flow method with specific similarity metrics to generate appearance-based models used to estimate the tracklet affinity between candidates. More recently, Iqbal et al. [4] and insafutdinov et al. [5], proposed similar approaches where the multi-person pose estimation and tracking are jointly modeled in a single formulation. They represent body joint detections for every frame of the sequence by a spatio-temporal graph and solve all plausible body pose trajectories for each person.

The main problem of these methods is that hand-crafted features are not sufficiently discriminative to identify faces with large appearance changes. Besides, some of these proposals do not consider the possibility of different camera shots, and/or do not exploit facial identification to enhance the identification of the people along the frames.

In recent years, Deep Learning techniques applied to facial recognition have led to a boost in the performance of multi-person re-identification. Parkhi et al. [6] proposed the Deep Face Recognition method where they use CNNs and the triplet-loss function to perform the recognition of faces along the frames. In [7], Jin et al. presented an end-to-end Face Detection method, where feature vectors for each face are extracted from the fc7 layer of the standard pre-trained CNN.
called Vgg-face (presented in [6]). It consists of vectors with 4,096 characteristics where two faces are compared through vector comparisons with higher relevance.

These methods achieve good results for face recognition, but their re-identification is lost when faces are occluded or hidden when turning over or faces present high variability due to different poses and orientations or illumination, shadows or blurring effects. How to correctly combine facial identification with extra semantic information is crucial to enhance the identification along the frames. Within this research, some authors have proposed a combination of facial and texture information to improve the tracking of people that appear in the video sequences. Lin et al. [8] developed a prior-less method for multi-face tracking. Facial feature vectors (obtained from the fc7 layer of the CNN presented in [6]) and simple clothes tracklets are linked from frame to frame to solve the tracking of the persons along the frames. The authors use Gaussian process to reduce the dimensionality without losing complex and important spatial-temporal information.

In a recent paper, Zhang et al. [9] proposed a multi-face tracking system for unconstrained videos. The authors used video-specific face representations using convolutional neural networks (CNNs) for recognizing faces along the sequences combined with the texture information obtained from the analysis of basic body regions.

These methods outperforms the re-identification results of previous ones thanks to the combination of facial analysis with the texture information of the people under tracking. Therefore, combining facial information with semantic characteristics such as texture information, 2D poses or type of clothes among others, is central to the multi-person re-identification problem.

In this work presented here, we address the re-identification problem by proposing a person re-identification method for unconstrained videos that matches this last group of proposals. We use CNNs to extract facial, texture and pose information of the people, and combine them to compare detected characters along the frames. The results demonstrate that our proposal achieves correct character tracking along the sequence, even in those frames where faces are occluded or characters present similar textures.

### III. Method

We propose a multi-person re-identification system for RGB videos where the frames are processed sequentially. Figure 1 shows the work-flow of the system. For each image, we compute three techniques based on CNNs to obtain reliable information from the images:

- Densepose [10], proposed by Guler et al., to obtain the correspondences between pixels belonging to the characters and the 3D model texture. By processing densepose at each frame, we can obtain a texture map of the detected people. For each pixel belonging to the character mask we obtain a vector in the (R, G, B, U, V, I) domain. (R, G, B) are the color channels of the pixel, (U, V) are the texture coordinates correspondence for the 3D model and (I) is the region of the body where the pixel belongs to. The body texture is defined for 24 body regions. We can see an example of these body regions in Figure 1.

- Instance-level human parsing, presented in [11] by Gong et al.. We use this semantic segmentation to detect all the characters that appear in the image and, for each character, we extract the head region, which will be used for facial recognition and to obtain the shape of the characters. Analogously to [8], we use the CNN VGG-face network [6] to obtain a vector of characteristics for each face. We extract the fully-convolutional layer fc7 of this network to have a 4,096 dimensional vector of characteristics which will be used to identify the face along the frames.

- 2D pose detection based on the Openpose method [12]. We extract 2D poses to exploit the pose similarity between characters detected from consecutive frames. Besides, we obtain their (x,y) coordinates of their position on the image to compute the matching decision according also to their proximity.

Once these techniques are computed, we have the following information for each detected character:

- Texture maps (τ).
- Spatial information (s): Bounding Box width, height and centroid.
- Facial identification (η): 4,096 dimensional vector to identify each face.
• 2D pose detection (φ), where we obtain the body joints position in (x,y) coordinates.

A. Work-flow description

The work-flow of the system for each frame is as follows:
1. At frame t, we have a database of tracked people (ψ) with N people being tracked.
2. We detect the number of people (K) present in the scene, by running the 2D pose detection [12]. We consider only the characters with a minimum area size, and with at least 5 joints detected. We call them candidates at frame t (β).
3. For each candidate k, we extract the texture, faces, poses and spatial parameters: βk = (τk, sk, ηk, φk).
4. We compute the Euclidean distance between K candidates (β) and N tracked persons (ψ) for each of the parameters τ, s, η, φ separately:
   \[ d_{k,n}(β_k, ψ_n) = (d_{τ_k,n}, d_{sk,n}, d_{η_k,n}, d_{φ_k,n}) \] (1)
   we normalize the distances so that
   \[ 0 ≤ (d_{τ_k,n}, d_{sk,n}, d_{η_k,n}, d_{φ_k,n}) ≤ 1. \] (2)
5. Once we have all \( d_{k,n}(β_k, ψ_n) \) distances, tracked persons (ψ) and candidates (β) match if the normalized distances accomplish one of these conditions:
   • Each \( (d_{τ_k,n}, d_{sk,n}, d_{η_k,n}, d_{φ_k,n}) \) is smaller than a defined threshold \( γ = 0.45 \).
   • If faces are very similar, even when textures present strong differences: \( d_{sk,n} < 0.2 \) and \( d_{φ_k,n} < 0.7 \) and \( (d_{sk,n}, d_{φ_k,n}) < 0.45 \).
   • If textures are very similar, but faces present differences: \( d_{τ_k,n} < 0.2 \) and \( d_{η_k,n} < 0.7 \) and \( (d_{τ_k,n}, d_{η_k,n}) < 0.45 \).

According to these conditions, if several candidates match one tracked person, we establish the final matching with the candidate \( β_k \) that presents the minimum average distance \( d_{k,n}(β_k, ψ_n) \).
6. With the matching correctly performed, we update the tracked people database with the characteristics of the associated candidate \( β_k = (τ_k, s_k, η_k, φ_k) \). This updating is performed by keeping 90% of the tracked character history and updating 10% with the new parameters.
7. Finally, candidates \( β_k \) non associated to any tracked person are considered as potential new characters to track in the successive frames. They will be accepted as person to track if their matching is performed for 5 successive frames.

B. Tracked People Texture Model

We use Densepose [10] to create the texture maps for each person (Figure 1 shows two examples). Hence, for each tracked person, we create a probabilistic model to characterize the texture map via pixel-wise Gaussian model [13] thus modeling each pixel of the texture with a Gaussian distribution in the RGB domain.

For each frame \( t \), we compute the distance between tracked characters and candidates by comparing one to one the probabilistic models of the tracked people \( τ_n \) with the texture maps of the candidates \( τ_k \). Therefore, we obtain the distance between the texture model and the candidate texture, by obtaining the percentage of non-matching pixels with respect to the total number of texture map pixels.

At each frame, the texture model of each tracked character is 10% updated with the texture of the associated candidate.

C. Camera Shots transitions Detection

Since our method deals with sequences with camera shot transitions, we detect these to apply a different processing using the method proposed in [14], which analyzes the color distribution of consecutive frames to detect strong variations. Hence, after a shot transition is detected, the continuity of the sequence has been broken and the spatial and 2D pose analysis are not reliable. Therefore, we perform the tracking of that frame considering only facial and texture analysis.

IV. RESULTS

The system has been tested on a number of different sequences with various people and scenarios. Examples of successfully solved conflicting situations are shown in Figure 2. The images show the tracked people with an id number and a coloured bounding box associated, which identifies the person along the frames. The first row of Figure 2 shows a sequence from a concert where one character disappears from the scene, and a new one appears in a posterior frame. Our system solves these shot changes, detecting already tracked characters, and generates new people to track that appear along the frames. Besides, the tracking is performed even when the people turn around and their faces are hidden, thanks to the clothes texture information used in the process. The second row shows an interview where two people are wearing similar clothes. In this situation, facial characteristics better discriminate the people and allow for their correct re-identification even though there are camera shot transitions. The third row shows a difficult sequence from an episode of The Big bang Theory TV sitcom, where four characters interact in the scene. This sequence presents camera shot transitions, partially detected faces and similar textures between people. The identification of the characters is correctly performed with an accurate association between tracked people and candidates at each frame.

Regarding the numerical evaluation, we assessed the quality of tracking people along the frames of our method using the weighted purity:

\[ W = \frac{1}{M} \sum_{n} (m_n \cdot p_n), \] (3)

where each person to track \( n \) contains \( m_n \) elements and its purity \( p_n \) is measured as the fraction of the largest number of appearances from the same person to \( m_n \), and \( M \) denotes the total number of appearances in the video. Table I shows the numerical comparison of the proposed algorithm with several reference methods. We use the Big Bang Theory database composed of the first five episodes from Season 1 of The Big Bang Theory TV Sitcom in order to compute our results.

As we can observe, our system outperforms existing reference methods, solving the tracking in difficult camera shots


**Fig. 2.** Qualitative results. Each row corresponds to a different sequence under evaluation. Every tracked person has a numerical index that denotes its correct tracking along the sequence.

| Methods         | S1-E01 | S1-E02 | S1-E03 | S1-E04 | S1-E05 |
|-----------------|--------|--------|--------|--------|--------|
| HOG [1]         | 0.37   | 0.31   | 0.37   | 0.36   | 0.29   |
| AlexNet [15]    | 0.47   | 0.31   | 0.45   | 0.36   | 0.29   |
| VGG-Face [6]    | 0.91   | 0.85   | 0.85   | 0.54   | 0.65   |
| SynTriplet [9]  | 0.95   | 0.95   | 0.93   | 0.84   | 0.86   |
| **Ours**        | 0.98   | 0.96   | 0.96   | 0.91   | 0.92   |

**TABLE I**

TRACKING RESULTS ON 5 BBT VIDEOS COMPUTING THE WEIGHTED PURITY.

transitions and when facial occlusions are present in the frames.

**V. CONCLUSIONS**

The tracking system that we have presented in this paper, performs an accurate multi-person identification in unconstrained video sequences that present camera shot transitions and strong characters variations along the frames. We use facial, texture, 2D pose and spatial analysis to perform the matching between tracked people and candidates for each frame. The results of the paper show how the system solves correctly some difficult situations where faces are partially occluded or people present similar textures.

**REFERENCES**

[1] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *IEEE conf. on computer vision and pattern recognition*, 2005, vol. 1, pp. 886–893.

[2] Xuemei Zhao, Dian Gong, and Gérard Medioni, “Tracking using motion patterns for very crowded scenes,” in *European conf. on computer vision*. Springer, 2012, pp. 315–328.

[3] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang, “Tracklet association with online target-specific metric learning,” in *IEEE conf. on computer vision and pattern recognition*, 2014, pp. 1234–1241.

[4] Umar Iqbal, Anton Milan, and Juergen Gall, “Posetrack: Joint multi-person pose estimation and tracking,” in *IEEE conf. on computer vision and pattern recognition*, 2017, pp. 2011–2020.

[5] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele, “Artrack: Articulated multi-person tracking in the wild,” in *IEEE conf. on computer vision and pattern recognition*, 2017, pp. 6457–6465.

[6] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, “Deep face recognition,” in *British machine vision association*, 2015, vol. 1, p. 6.

[7] SouYoung Jin, Hang Su, Chris Stauffer, and Erik Learned-Miller, “End-to-end face detection and cast grouping in movies using erdos-renyi clustering,” in *IEEE int. conference on computer vision*, 2017, pp. 5276–5285.

[8] Chung-Ching Lin and Ying Hung, “A prior-less method for multi-face tracking in unconstrained videos,” in *IEEE conf. on computer vision and pattern recognition*, 2018, pp. 538–547.

[9] Shun Zhang, Jia-Bin Huang, Jongwoo Lim, Yihong Gong, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang, “Tracking persons-of-interest via unsupervised representation adaptation.” *Int. journal of computer vision*, vol. 128, no. 1, pp. 96–120, 2020.

[10] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *IEEE conf. on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[11] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin, “Instance-level human parsing via part grouping network,” in *European conf. on computer vision*, 2018, pp. 770–785.

[12] Zhe Cao, Tomas Simon, Shah-En Wei, and Yaser Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields, in *IEEE conf. on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[13] Chris Stauffer and W Eric L Grimson, “Adaptive background mixture models for real-time tracking,” in *IEEE conf. on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252.

[14] Dario Cazzato, Marco Leo, Pierluigi Carcagni, Cosimo Distante, Javier Lorenzo-Navarro, and Holger Voos, “Video indexing using face appearance and shot transition detection,” in *IEEE int. conference on computer vision workshops*, 2019.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.