People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation

Kheir-Eddine AZIZ, Djamel MERAD, and Bernard FERTIL
LSIS - UMR CNRS 6168
163, Avenue of Luminy, 13288 Marseille Cedex 9, France
kheir-eddine.aziz@univmed.fr, djamal.merad@univmed.fr, bernard.fertil@univmed.fr

Abstract

In this paper, we present a new person re-identification method based on appearance classification and silhouette part segmentation. In crowded areas, heads are considered as most apparent parts, hence the typical advantage of using the skeleton graph for the head detection and location of people after partial occlusion. The appearance classification consists in characterizing the appearance of a person into two classes, the frontal and the back appearance, using head detector and the orthogonal iteration algorithm for head pose estimation. The silhouette part segmentation divides the silhouette into three horizontal parts, ideally corresponding to head, torso and legs using skeleton graph and head detector. Our approach is robust to real world situations, in particular to variations in scales, human pose, illumination and clothes appearance changes. It also allows to reduce the confusion cases among people appearance and the amount of falsely matches.

1. Introduction

Person re-identification is a crucial issue in multi-camera tracking scenarios, where cameras with non-overlapping views are employed. Considering a single camera, the tracking captures several instances of the same individual, providing a volume of frames. The re-identification consists in matching different volumes of the same individual, coming from different cameras. In the literature, the re-identification methods that focus solely on the appearance of the body are dubbed appearance-based methods, and can be grouped in two sets. The first group is composed by the single-shot methods, that model a person analyzing a single image [5][18]. They are applied when tracking information is absent. The second group encloses the multiple-shot approaches; they employ multiple images of a person (usually obtained via tracking) to build a signature [2][4][6][12]. In [2], each person is subdivided into a set of horizontal stripes. The signature is built by the median color value of each stripe accumulated over different frames. A matching between decomposable triangulated graphs, capturing the spatial distribution of local temporal descriptions, is presented in [4]. In [6], a signature composed by a set of SURF interest points, collected over short video sequences, is employed. In [12], each person is described by local and global features, which are fed into a multi-class SVM for recognition. Other approaches simplify the problem by adding temporal reasoning on the spatial layout of the monitored environment, in order to prune the candidate set to be matched [7], but these cannot be considered purely appearance-based approaches.

In this paper, we present a novel multiple-shot person re-identification method based on appearance classification and silhouette part segmentation. The first step consists in classifying the people appearance into two classes, frontal and back appearance, using head detector and a head-camera distance. This classification provides robustness against illumination and clothes appearance changes (see Figure 1. Examples of matching of interest points for a given person seen under 2 same viewpoints (better matching with appearance classification)(a, c) and from a different viewpoints (b)).

The second step consists in segmenting the silhouette of the person in different horizontal parts (head, torso and legs) for re-identification. This segmentation is based on the head detector and skeleton graph. This step is integrated in the single-camera tracking phase. Then, the complementary signatures are extracted of each part of the silhouette: SIFT, SURF and Spin image interest points. For the matching,
we use the cross-bin metric based Earth Mover’s Distance (EMD) variant to associate the regions of one silhouette to another. This segmentation is robust to variations in scales (see Figure 2).

Figure 2. The matching results for two given person with/without segmentation. (a) Shows a false matching example, (b) shows a correct matching example. Always, the corresponding horizontal descriptor versions reduce the number of false matches. (c) Shows an ideal matching in the scale change situation.

2. The proposed method

The proposed system is illustrated in the Figure 3. The person re-identification system.

2.1. Head detection and tracking

To compute foreground masks we use a forward-backward approach developed by Ge et al.[3]. The forward-backward background subtraction pass will produce clean foreground masks for times less than T. This method (off-line) is feasible in a real-time system using delayed online process. It consists to introduce a short time delay. In this case, online systems can be like off-line methods with access to a small amount of future information.

For each detected region (individual/group human), we compute the graph skeleton for each of them using the approach developed by Thome et al.[13]. Unlike the methods from discrete geometry (distance transform, etc.), which are sensitive to noise, the algorithm that we used is an algorithm of computational geometry (Delaunay graph). Computational geometry’s algorithms are more robust than discrete geometry’s, when the object is invariant under the most important geometrical transformation. The filtering and smoothing operations provide a skeleton which is more robust and homotopic to the silhouette [11] as shown in Figure 4.

In the crowded area, heads are considered as the most apparent parts of the skeleton. At this step, to detect the heads we adopt the algorithm used in [11]. We are only focused on the points’ set having a single neighbor in the skeleton graph. The segment corresponding to the extreme node is subsequently taken and its degree inclination compared to the vertical axis is calculated. If the degree tilt is within $[-\theta, \theta]$, the segment is classified as the head of person [11].

After detection phase, we use the particle filter [17][14] to track each detected head.

2.2. The appearance classification

The single-camera tracking output usually consists in a sequence of consecutive images of each individual head in the scene. The appearance of the person is reached by calculating the distance between the tracked head and the camera. If this distance increases we are talking about frontal pose, otherwise it’s about back pose (see Figure 5).

To calculate this distance, we suppose the head space coordinates model $\{X_h, Y_h, Z_h\}$ and the model of head. The four 2D corners of the head model in the head space coordinates model are called $A$, $B$, $C$ and $D$ as shown in Figure 6.

This distance consists in finding the rigid transformation $(R, T)$ from $C$ to $H$ (Equation 1).

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = [R|T] \begin{bmatrix} X_h \\ Y_h \\ Z_h \end{bmatrix}$$

We use the method named Orthogonal Iteration (OI) algorithm, proposed by Lu et al.[10]. To estimate the head
pose, this algorithm used an appropriate error function defined in the local reference model of head. The error function is rewritten to accept the iteration based on the classical solution of the 3D pose estimation problem, called absolute orientation problem. This algorithm gives exact results and converges quickly enough.

![Figure 6. Different appearance of one person.](image)

If the component $T_z$ of the translation vector $T$ increases (see Figure)

**Algorithm 1** Distance head-camera for one person

**Require:** Internal parameters of the calibrated camera

**Require:** `heads_list` : list of the tracked head for one person

**Require:** $[X_h, Y_h, Z_h]$ = [(0, 20, 0); (20, 20, 0); (20, 0, 0); (0, 0, 0)] : The coordinates of the head model in the head space coordinates model

**Require:** $[a, b, c, d]$ : The four corners of the detected head in the image plane

1: for $i$ = 0 to `heads_list.size` do
2: $[R_i, T_i] \leftarrow$ OI($[X_i^c, Y_i^c, Z_i^c]$, $[X_h, Y_h, Z_h]$)
3: $[R_i+1, T_i+1] \leftarrow$ OI($[X_i^c+1, Y_i^c+1, Z_i^c+1]$, $[X_h, Y_h, Z_h]$)
4: if $T_{iz} > T_{iz+1}$ then
5: `Appearance_Person_i` $\leftarrow$ Frontalpose
6: else
7: `Appearance_Person_i` $\leftarrow$ Backpose
8: end if
9: end for

2.3. The torso and legs segmentation

After the head detection, it’s easier now to locate the silhouette and divide it into three parts (head, torso and legs) in the single and partial occlusion situation. For that, we re-project a body-model on each detected head in the image plane to get the body of the person. We suppose the average width and height of one person in world space coordinate is respectively 40cm and 170cm. when the height of the head is 20cm, the torso is 70cm and the legs are 80cm (see Figure.

In the case of partial occlusion, the body of one or more persons are often hidden by the bodies of the others. For two persons for example, we can separate them with the head detector (see Figure.

![Figure 7. The silhouette part segmentation. (a) Original image, (b) Head detection, (c) torso and legs segmentation.](image)

Figure 7. The silhouette part segmentation. (a) Original image, (b) Head detection, (c) torso and legs segmentation.

![Figure 8. The advantage of the body-model re-projection. (a) Original image, (b) foreground mask, (c) skeleton graph, (d) torso and legs segmentation after head detection and re-projection of the body-model.](image)

Figure 8. The advantage of the body-model re-projection. (a) Original image, (b) foreground mask, (c) skeleton graph, (d) torso and legs segmentation after head detection and re-projection of the body-model.

In order to accomplish, we calculate the spatial distance among the heads. If this distance is below a threshold, we keep only the head close to the camera and we associate the body to this head (Algorithm

2.4. Descriptors and features matching

In the following we briefly explain the SIFT [9], SURF [1] and Spin image [8] descriptors which offer scale and rotation invariant properties.

SIFT (Scale Invariant Feature Transform) descriptors are computed for normalized image patches with the code provided by Lowe [9]. A descriptor is a 3D histogram of gradient location and orientation, where location is quantized into a $4 \times 4$ location grid and the gradient angle is quantized into eight orientations. The resulting descriptor is of dimension 128. Each orientation plane represents the gradient magnitude corresponding to a given orientation. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

SURF (Speeded Up Robust Features) is conceptually
similar to the SIFT descriptor, the 64-dimensional SURF descriptor also focuses on the spatial distribution of gradient information within the interest point neighborhood, where the interest points itself can be localized by interest point detection approaches or in a regular grid. The SURF descriptor is invariant to rotation, scale, brightness and, after reduction to unit length, contrast.

Spin image is a histogram of quantized pixel locations and intensity values [8]. The intensity of a normalized patch is quantized into 10 bins. A 10 bin normalized histogram is computed for each of five rings centered on the region. The dimension of the spin descriptor is 50.

For the matching, we use the cross-bin measure two compare between two descriptors. In [15], Pele et al. proposed a linear time algorithm for the computation of the EMD variant, with a robust ground distance for oriented gradients.

Given two histograms P, Q; the EMD as defined by Pele et al. [15] is:

\[ \text{EMD}(P, Q) = \min_{\pi} \sum d_{ij} \pi_{ij} \text{subject}\] 

\[ \text{EMD}(P, Q) = \sum_{i,j}(-\alpha \max_{i,j} d_{ij}) \text{subject}\] 

where \( f_{ij} \) denote the flows. Each \( f_{ij} \) represents the amount transported from the \( i^{th} \) supply to the \( j^{th} \) demand. \( D_{ij} \) denote the ground distance between bin \( i \) and bin \( j \) in the histograms. EMD has two advantages over Rubner’s EMD [16] for comparing SIFT descriptors. First, the difference in total gradient magnitude between SIFT spatial cells is an important distinctive cue. Using Rubner’s definition this cue is ignored. Second, EMD is a metric even for non-normalized histograms. We use the same metric to match the SIFT, SURF and Spin image descriptors.

3. Experimental results

3.1. Re-identification method

Re-identification consists in associating each person of set \( B \) to the corresponding person of set \( A \). This association depends on the content of tree sets: 1) single-shot vs single-shot, if each image represents a different individual; 2) multiple-shot vs single-shot, if each image in \( B \) represents a different individual, and in \( A \) a single person is described by a multiple images signature; 3) multiple-shot vs multiple-shot, if both \( A \) and \( B \) contain signatures from multiple images (our case).

In our re-identification method, the association of each person \( P_i \) to the corresponding person \( P_j \) is done by a voting approach: every interest point extracted from the set \( P_i \) is compared to all models points of the person \( P_j \), and a vote is added for each model containing a close enough descriptor. We just match the interest points of the same regions (head vs head, torso vs torso, legs vs legs) and the same class of appearance (back vs back and frontal vs frontal). Eventually, the identification is made with the highest voted for the model. Two descriptors are matched, if the first descriptor is the nearest neighbor to the second and the distance ratio between the first and the second nearest neighbor is below a threshold. In our knowledge, there is no available benchmark for evaluation of person re-identification based on multi-shot appearance with non-overlapping camera. We therefore decided to conduct a first experimental evaluation of our proposed method on an available series of videos showing persons recorded with three non-overlapping camera. These videos include images of the same 9 person seen by three non-overlapping camera with different viewpoints (see Figure).

3.2. Evaluation of the method

To evaluate the overall performance of our proposed method, we select the set of person detected in the camera one as test data (request) and the set of the person detected in the other cameras as reference data (model). The model for each person was built with 8 images (4 images for the frontal appearance and 4 images for the back appearance collected during tracking phase). Likewise, for a test, each query was built with 8 images.

The re-identification performance evaluation is done with the precision-recall metric:

\[ \text{Precision} = \frac{\text{correct matches}}{\text{correct matches} + \text{false matches}} \]

\[ \text{Recall} = \frac{\text{correct matches}}{\text{query matches}} \]

Algorithm 2 Association Head Body

| Require: heads_list : list of the tracked heads |
| 1: for \( i = 0 \) to heads_list.size do |
| 2: for \( j = i + 1 \) to heads_list.size do |
| 3: dist ← dist(centroid(head_i), centroid(head_j)) |
| 4: if dist < threshold then |
| 5: partial occlusion state ← true |
| 6: dist_head_i_camera ← dist(head_i, camera) |
| 7: dist_head_j_camera ← dist(head_j, camera) |
| 8: if dist_head_i_camera < dist_head_j_camera then |
| 9: \{Person j covered by person i\} |
| 10: silhouette ← Associate(body, head_i) |
| 11: head ← head_i |
| 12: else |
| 13: \{Person i covered by person j\}; |
| 14: silhouette ← Associate(body, head_j) |
| 15: head ← head_j |
| 16: end if |
| 17: [torso, legs] ← seg(silhouette) |
| 18: end if |
| 19: end for |
| 20: end for |
4. Conclusion

In this paper, we proposed a new person re-identification based on the appearance classification and silhouette part segmentation. Firstly, we classified the people into two appearance classes by calculating the geometric distance among heads and the camera. Based on the head detector by skeleton graph, it is easy to calculate this distance in crowded environment. The head detection can also help to locate people after partial occlusion using body-model. Secondly, we segment the silhouette into three horizontal parts: head, torso and legs using skeleton graph and head detector. Employing the classification, we obtained the novel best performances to reduce the illumination variations and clothes changes. The silhouette part segmentation decrease the false matches and the confusion among the silhouettes that have similar overall appearance and different scale. The advantage of combining both methods improve the re-identification performance. As future work we plan to extend our method by fusion of multiple signatures and using a decision tree for the re-identification.

References

Figure 12. The precision vs recall curves for person re-identification experiment: (a) with appearance classification. (b) with silhouette part segmentation, (c) with appearance classification and silhouette part segmentation.

