View-invariant Person Re-identification with an Implicit Shape Model

Kai Jüngling
Fraunhofer IOSB
Ettlingen, Germany
kai.juengling@iosb.fraunhofer.de

Michael Arens
Fraunhofer IOSB
Ettlingen Germany
michael.arens@iosb.fraunhofer.de

Abstract

In this paper, we approach the task of appearance based person re-identification for scenarios where no biometric features can be used. For that, we build on a person re-identification approach that uses the Implicit Shape Model (ISM) and SIFT features for re-identification. This approach builds identity models of persons during tracking and employs these models for re-identification. We apply this re-identification, which was until now only evaluated in the infrared spectrum, to data acquired in the visible spectrum. Furthermore we evaluate view independence of the re-identification approach and introduce methods that extend view invariance. Specifically, we (i) propose a method for online view-determination of a tracked person, (ii) use the online view-determination to generate view specific identity models of persons which increase model distinctiveness in re-identification, and (iii) introduce a method to convert identity models between views to increase view independence.

1. Introduction

Object, and more specifically person tracking is an indispensable part of many of today's computer vision applications. In many cases where high-level video analysis is necessary, specifically in areas like visual surveillance, it is not sufficient to track a person while it continuously appears in the field of view of a single camera, but to re-identify a person after it has left the systems field of view and reenters it again. The system's field of view hereby can refer to a single or multiple cameras. Even when referring to a system with a single camera, re-identification can be necessary, e.g., to determine if a person visits a shop-window multiple times, to determine if the same or a different person picks up a bag that someone has left before, and also to detect suspicious behavior which can be constituted by visiting the same place multiple times.

In this paper, we approach the problem of person re-identification. We build on the re-identification approach described in [8] and revised in [9]. The essential of this approach is, that it builds on the Implicit Shape Model (ISM) [10] and SIFT [11] features for both person tracking and re-identification. The re-identification uses the SIFT features and ISM characteristics collected online during the tracking of a person to re-identify this person later on in a multi-staged approach. The overall approach has several advantages over state-of-the-art re-identification techniques:

(i) By employing only SIFT features for detection, tracking and re-identification, the proposed system is most independent of the employed sensor. Unlike most other re-identification approaches [6, 4, 15], the ISM re-identification does not employ sensor specific features like color, which makes it applicable for the case of data acquired in the visible and infrared spectrum.

(ii) The multi-stage approach with increasing computational cost allows for very efficient re-identification since the computational cheap first stages can be used to reduce the amount of data (candidate models from the database) that has to be considered in the last stage.

(iii) Compared to most other state-of-the-art approaches like [3, 4], this approach is applicable in real applications since it is integrated with a detection and tracking strategy. Specifically, it does not rely on manual annotation of people like [3, 4] and builds models for re-identification online without an offline training step like [2, 5].

We extend this re-identification approach in multiple ways: (i) we introduce a method for online view determination of a tracked person, (ii) we use the online view determination for the generation of view specific models which increase distinctiveness of models in re-identification, and (iii) we introduce a method to convert a model from one view to another on the basis of the ISM and SIFT. In addition, we apply the re-identification approach to the task of person re-identification in the visible spectrum. In section 2, we give an overview of the principal re-identification approach. Section 3 discusses view independence and introduces methods for view determination and identity model transformation. Section 4 presents an evaluation and section 5 concludes.

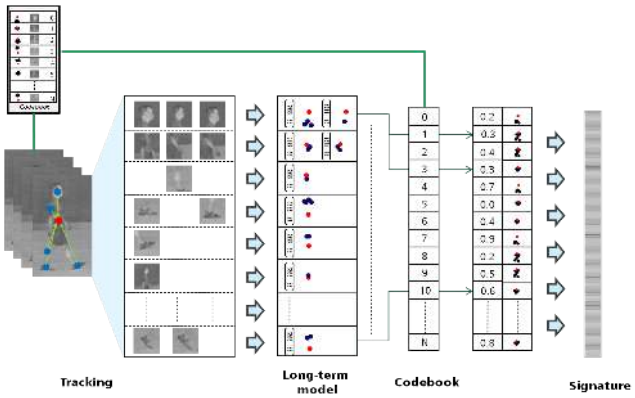

Figure 1. Generation of person identity models during tracking: SIFT features found on a person during tracking are stored in a long-term feature model. In addition to the high-dimensional feature model, codebook activation signatures which allow for fast model matching, are stored.

Figure 2. Overview of the three-staged ISM re-identification approach.

2. Person re-identification

We build on the re-identification approach described in [8] and revised in [9]. Briefly, this re-identification builds on the ISM and SIFT features and is closely coupled with an ISM based person detection and tracking approach making it applicable in real-world applications where tracking and re-identification is always a coupled problem since a good tracking performance is prerequisite for good re-identification performance. The overall approach has three main components. *Person detection* uses an Implicit Shape Model based object detection approach to detect people in input imagery. Since a dedicated object detector and no foreground or motion detection is used, no assumptions on stationary camera or application scenario are made. These characteristics are retained in *Tracking*, where the task is to build trajectories of persons while they continuously appear in the field of view of a single camera. The re-identification builds on tracking by using the SIFT feature models collected during tracking. This circumstance is visualized in figure 1. Here, SIFT features found on a person during tracking are visualized by the according image patches. These features are integrated into a *long-term identity model* containing all appearance variations which occur during tracking. Together with the SIFT descriptors, which are clustered to reduce quantity of data, the spatial feature distribution, which describes the position where certain features have been found on the person relatively to the object center, is recorded. Together with this feature model, an ISM activation signature is stored. The calculation of this activation signature is shown in the right part of figure 1. Here, the codebook (this codebook represents the general appearance of the class "person" by a number of SIFT prototypes that have been generated in a training step) which was used for object detection (see [7] for details) is used to build the signature. For each feature in the long-term model, the codebook prototype affiliation is determined by matching the feature to the codebook (this has already been performed in object detection). This is in a way a description of the person in terms of a histogram, which in this context, is commonly referred to as a "bag of words" representation. The key difference to a usual "bag of words" representation is that here, information of a time period is integrated into a single model, and, that in addition to the "word affiliation", the spatial distribution for each word is stored. So basically, the signature encodes which codebook entries are activated by a person to what extend. Person re-identification is performed on the basis of this information in a three staged approach which is outlined in figure 2. In the first stage, the low-dimensional codebook activation signature is used for person description and matching. By that, fast matching of a query model with a database of persons is possible because only low dimensional vectors (codebook dimension N, which is usually between 200 and 1000) are to be compared. In the second stage, the spatial distribution of the codebook signature is added to the comparison. This increases matching complexity only slightly – for each codebook signature entry, the spatial distribution has about 10 2D entries, which adds $N * 2 * 10^2$ comparison in relation to stage 1 – while distinctiveness is increased strongly since in addition to encoding of appearance, now, structural aspects are encoded too. In stage 3, the high dimensional SIFT feature models are compared. This increases the computational demand on this stage because here, sets of 128-dimensional SIFT descriptors are to be matched. The important point is that distinctiveness is increased significantly compared to the former stages. A possible way to deal with the high computational demand for comparison on stage 3 is to use the three stages in a cascade to reduce the amount of data that is to be matched in subsequent stages. This is depicted in figure 2.

3. Viewpoint invariance of person re-identification

The re-identification approach described in the last paragraph has some shortcomings. Since it uses SIFT features together with a structural modeling by the ISM, it is not inherently view independent like approaches that model person appearance by globally valid features, e.g. a person-global color histogram. Although the global validity of such appearance descriptions is not necessarily true in every case (e.g. not every person looks the same from front and side view), the assumption that such a color histogram is a good cue for view independent description of person appearance is valid in most cases. This is not true in this approach, where SIFT features are used for person appearance description, because these are (i) only view independent to a certain degree, and (ii) since they model texture found on a person and shape of a person, we cannot assume that these features are transferable between different views of a person (e.g. texture found on a person's front side is typically not the same as texture found on a person's back). Since the ISM-SIFT appearance modeling has several advantages over appearance description by simple color histograms (e.g. higher distinctiveness, more detailed modeling, applicability in monochromatic imagery, applicability in other modalities like infrared) it is desirable to keep these advantages but at the same time eradicate the shortcomings. For that reason, we introduce an extension of the ISM person re-identification which allows for view dependent generation of person models and conversion of different views into each other. This has the two main advantages that discriminate power and view independence is increased.

3.1. Viewpoint classification and selection

In ISM re-identification, SIFT features found on a person during tracking are stored in a single model for each tracked person. Since a person might be visible from different viewpoints during tracking (due to movement of the person in the scene or due to camera motion), this model potentially includes appearance information from different viewpoints. This means that two models of completely different views might be compared in person re-identification. This is not desirable since (i) the same person might look very different from different views and (ii) different persons might accidentally have similar appearance in different views. For that reason, it is desirable to store view specific models during tracking and use them in person re-identification.

Thus, in a first step, the current view of a tracked person is to be determined (considering a discrete set of possible views). For that, we can use the short-term SIFT models from tracking which encode the current appearance (the appearance of the recent history) of a person and thereby do not only hold the identity of a person, but also, in parts, the current view. To determine the actual view of a person, these models are compared to general, view-specific models which were generated offline on a training set. The training set must include samples from different persons for each view, so that identity information can be discarded and only the view information can be retained. In practice this is done by clustering the feature models of the same view from different persons and discarding features which are not seen often enough in training data and thus can be assumed to contain identity specific information. The view of a tracked person is then determined by matching the feature model to all view-models and picking the best matching view. This is a simple approach but works quite well for our scenario. The drawback of this approach is its high computational demand – the high dimensional feature models have to be matched at every instant of time. Thus, this approach might be replaced by a classification approach that works directly on the ISM basis without the need for an additional matching step.

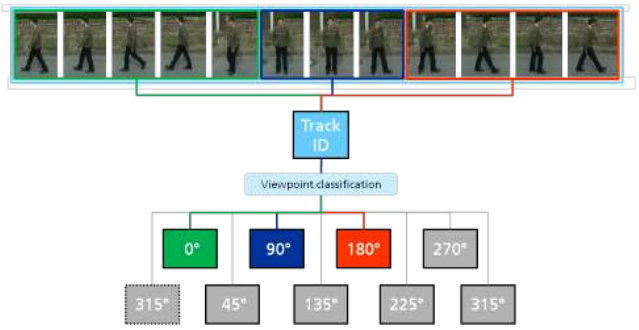

Figure 3. View classification during tracking and generation of view-specific person models.

Knowledge about the view of a tracked person can be used to store view-specific person models as depicted in figure 3. SIFT features currently found on the tracked person are stored in a view specific model that refers to the view which was determined for the currently tracked person. The view thereby is picked out of a discrete set of possible and previously defined views.

In re-identification, based on the view of a currently tracked person, we can now pick the correct view from the database for comparison. This approach increases distinctiveness but is only helpful if the relevant view (or a similar view) is available for all database models. Since this might not always be the case in reality, it is necessary to be able to compare two different views to each other. While in cases where views differ due to object inherent differences between views (e.g. front and side view of a person), it might not be possible (or at least not helpful) to convert views into each other, in other cases where object appearance can expected to be similar but appearance description does not in-

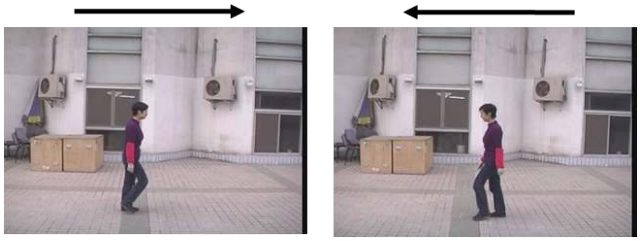

Figure 4. Example case where ISM mirroring is helpful.

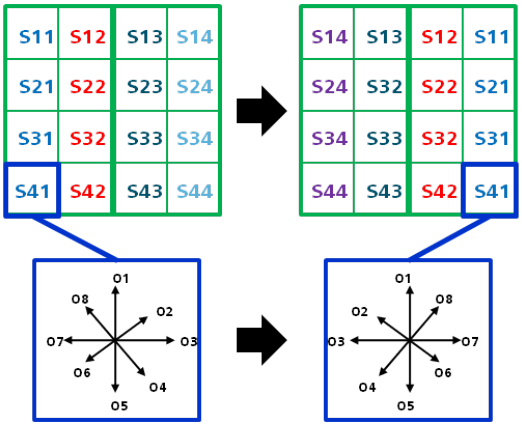

Figure 5. Mirror transformation of the SIFT descriptor.

clude the invariance, conversion between views might be useful. Such a case is visibility of persons from opposing sides. This happens when a person first moves through the camera's field of view from left to right and a second time from right to left (in reality: passing a corridor on the way there and on the way back) as shown in figure 4. This is also relevant in multi-camera networks where cameras are often mounted opposingly. Due to symmetry between left and right side of a person, which can commonly be expected, the opposing side views of a person can be transformed into each other by mirroring. So, this case which is often relevant in practice can be treated by a mirror transformation. This transformation is introduced in the next section for the ISM re-identification.

3.2. Identity model transformation

A case where the mirror-transformation motivated in the last section is relevant is shown in figure 4. Here, a person is visible from opposing sides. The models generated in these two views can be converted into each other by a mirror transformation.

To perform the mirror transformation for an identity model, both the ISM and the SIFT features are to be transformed. SIFT transformation is performed as shown in figure 5 (see [12]). Both, the spatial histogram (top) and the orientation histograms inside the spatial histogram (bottom) are transformed. The spatial histogram can be simply mirrored using the middle vertical axis. The orientation histograms are mirrored using O1 and O5 as mirroring axis. An important point is that SIFT transformation can be performed directly on basis of the descriptor vector without recalculation of the descriptor itself. This is very important because the descriptor calculation is very computational expensive. In fact, no explicit transformation has to be performed because mirroring can be performed by alternating descriptor entries during matching.

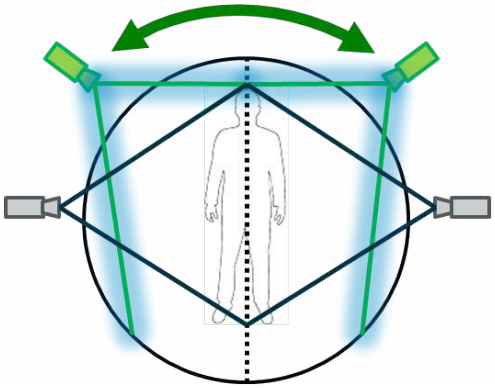

Figure 6. Symmetry that can be assumed for a person.

Since re-identification on all levels involves the features' codebook activation, this activation has to be mirrored too. Unfortunately, this can only be done by matching all features to the codebook using the mirror transformation during matching. Since this is an tremendous matching overhead, in future, mirroring of codebook activations should be performed in a more efficient way. One possible approach might be to calculate, for every codebook prototype, the mirror prototypes, so that a "normal" activation can be transformed directly into a mirror-activation without additional matching. Another possible approach might be to integrate multiple codebooks into a multi-view model as proposed in [14]. Using this mirror transformation, the two hemispheres of the view-sphere shown in figure 6 can be transformed into each other. Thus, 50% of possible view-transformations are covered with one model-transformation. Remaining views are more similar, which means, transformations between remaining views might be covered by inherent ISM and SIFT view invariance without the necessity of an explicit transformation. Regarding ISM, view dependence depends only on the similarity demand for spatial distributions. If this demand is relaxed, view independence can be increased. It is important to note that discriminative power decreases when view independence is increased. SIFT features are viewpoint invariant to a certain degree. Lowe [11] states, that in his experiments, he gets matches for 50% of the features when the viewpoint is changed by $50°$. In the ISM, the location of a feature is encoded by the center offset. Thus, viewpoint invariance

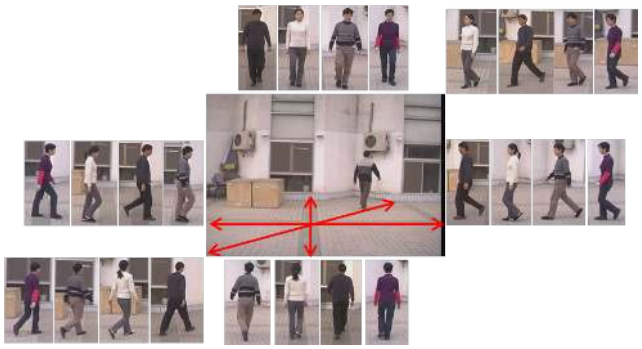

Figure 7. The CASIA A dataset.

of the ISM (when assuming correct matches) is determined by the variance that is allowed for the center offsets. In re-identification this is determined by the maximum deviation that is allowed for the offsets between two person models.

4. Evaluation

Evaluation is performed on the CASIA A dataset [1]. Here, we compare our ISM re-identification approach to other approaches and evaluate view-independence of ISM re-identification in detail. Sample images of the scene and the persons are shown in figure 7. Person size varies from 20x70 to 50x150 pixels. This dataset is well suited for evaluation of view-independence because it includes video sequences of persons which move through a scene from 6 different directions ($0°, 90°, 135°, 180°, 270°, 315°$). For every person in the dataset, two sequences are available for each walking direction. Thus a rating of the re-identification view-independence is possible using this dataset. For that, every person in the dataset (16 persons) is tracked in all available views to build the instance models. For 16 persons and 6 views this makes a total of 96 different view-person combinations. For each person, two sequences are available. Model building in these sequences results in 2 instance models per person and view. These are used as query and database model respectively. Since the distance measure we use for re-identification is not symmetric, each model serves as query and as database model.

Evaluation is carried out using the same performance measures as in [8]. Since in this paper, we do not perform an open-set evaluation, the relevant rate is the *Correct Classification Rate (CCR)* which is the ratio of correct classifications and models in the database when performing queries for all persons in the database.

For evaluation, the database is filled with models of a single view of every person. Models of a second view serve as query models. Every model of that view is tested against the database models. E.g., the database contains models of view $90°$, queries are made with models of view $180°$. This evaluation is performed for every possible database-view query-view combination. By that, we gain an exact assessment of re-identification view-independence.

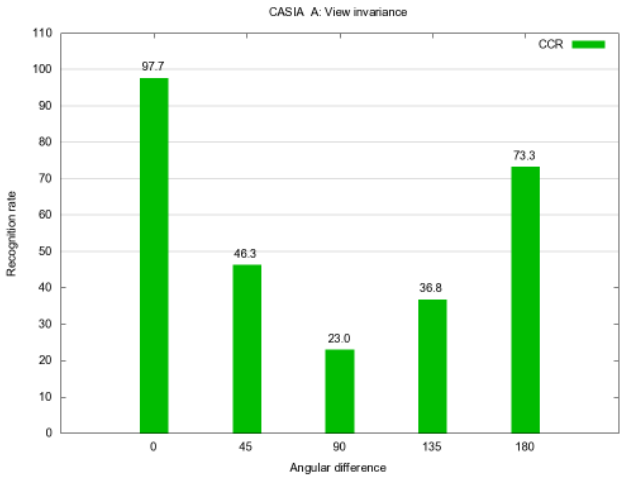

Figure 8. Re-identification results of the CASIA A dataset. Bars show CCR for different transformation angles.

Table 1. Correct Classification Rate for different viewpoint-combination of the CASIA A dataset.

Angle	**0**	**90**	**135**	**180**	**270**	**315**
0	93	25	42	81	27	57
90		100	36	20	67	34
135			100	50	36	72
180				93	20	25
270					100	42
315						100

Results of this scenario are shown in figure 8 and table 1. Figure 8 shows re-identification performance at different view transformation angles. Table 1 shows the detailed results for the different view combinations. Here, $0°$ refers to a person moving from right to left orthogonal to the camera axis, $90°$ refers to a person walking away from the camera, parallel to the cameras axis and so on.

As we can see from the chart in figure 8, best performance is of course reached when query and database view are the same. Here we gain a nearly perfect performance with a CCR of 97.7%. As we see in the detailed results in table 1, the performance is best when either part of the back or part of the front is visible. In both side views ($0°$ and $180°$), not all query models are classified correctly.

Performance decreases with increasing view difference. At a transformation angle of $45°$, 46.3% of the query models are classified as the correct database model. Performance decreases to 23.0% at a transformation angle of $90°$. Since this performance is rather bad, this could be a possible case, where the system could report that no conclusive decision could be made due to the lack of information - this

Table 2. Correct Classification Rate of different re-identification approaches on the CASIA A dataset when considering the same view for query and database model.

NN+NED	NN+STC	ENN+NED	HTI	ISM
62.1	68.75	83.3	94.6	97.7

is reasonable since we know that no appearance correlation exists between the front and side view of a person (when using local features to model the appearance). Performance increases again at angles of $135°$ and $180°$. This is due to the use of the mirror transformation which is employed to convert models in case of angles of $135°$ and $180°$. As we see from table 1, this leads to a good performance of 81% in the specific case where persons are visible from side-view $(0 - 180°)$. General performance in case of a $180°$ angle is 73.3%. This includes cases where persons are visible from front and back. Although visual appearance (in terms of texture found on a person) is not typically similar for a person in front and back view, re-identification performance is rather good with 67% ($90°$ and $270°$) in this case. This is because ISM re-identification does not only model texture found on a person but also the shape of a person which is very similar for front and back view of a person.

To assess the quality of ISM re-identification, we compare to other (gait recognition based) re-identification approaches which have been evaluated on the same dataset [16, 13]. Results of these approaches, which perform re-identification using gait-recognition, are shown in table 2. Comparison is only possible for the same view because the other approaches do not consider different views. As we see, with 97.7% for same-view re-identification, our approach clearly outperforms the other approaches on this dataset.

5. Conclusion

In this paper, we presented an extension to an Implicit Shape Model and SIFT based person re-identification approach that increases re-identification distinctiveness and view independence. We showed that this extended re-identification approach can be successfully employed for re-identification in the visible spectrum under view variations, while the original approach was formerly only applied to data from the infrared spectrum that did not contain view variations. Performance comparison showed that ISM re-identification outperforms other re-identification approaches on this dataset.

References

[1] Casia gait database, http://www.sinobiometrics.com. obtained from http://www.cbsr.ia.ac.cn/english/gait

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using haar-based and dcd-based signature. In *Proc. Advanced Visual and Signal based Surveillance*, pages 1528–1535, 2010.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. Advanced Video and Signal based Surveillance*, pages 435–440, 2010.

[4] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006.

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. European Conference on Computer Vision*, pages 262–275, 2008.

[6] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. Computer Vision and Pattern Recognition*, pages 26–33, 2005.

[7] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. In *Proc. Conference on Computer Vision (ICCV Workshops)*, pages 1129–1136, 2009.

[8] K. Jüngling and M. Arens. Local feature based person reidentification in infrared image sequences. In *Proc. Conference on Advanced Video and Signal based Surveillance*, pages 448–454, 2010.

[9] K. Jüngling and M. Arens. A multi-staged system for efficient visual person reidentification. In *Proc. of the Conference on Machine Vision Applications*, pages 1–8, 2011.

[10] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] R. Ma, J. Chen, and Z. Su. Mi-sift: mirror and inversion invariant generalization for sift descriptor. In *Proc. ACM Image and Video Retrieval*, pages 228–235, New York, NY, USA, 2010. ACM.

[13] D. Tan, K. Huang, S. Yu, and T. Tan. Efficient night gait recognition based on template matching. In *International Conference on Pattern Recognition*, volume 3, pages 1000–1003, 2006.

[14] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool. Towards multi-view object class detection. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, June 2006.

[15] D. Truong Cong, L. Khoudour, C. Achard, and L. Douadi. People detection and re-identification in complex environments. *IEICE Transactions on Information and Systems*, 93:1761–1772, 2010.

[16] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 25:1505–1518, 2003.