Improving Face Recognition in Low Quality Video Sequences: Single Frame vs Multi-frame Super-Resolution

Andrea Apicella, Francesco Isgrò, and Daniel Riccio(✉)

Università degli Studi di Napoli Federico II, Naples, Italy
and.api87@gmail.com, {francesco.isgro,daniel.riccio}@unina.it

Abstract. Re-Identification aims to detect the presence of a subject spotted in one video in other videos. Traditional methods use information extracted from single frames like color, clothes, etc. A sequence in time domain of consecutive subject images could contain a greater amount of information compared with a single image of the same subject. Typically, these sequences are taken from surveillance cameras at very poor resolution. Even with modern cameras the resolution can be a problem when dealing with a subject who is far from the camera. A possible way of handling low resolution images is by using a multi-frame super-resolution algorithm. Multi-frame super-resolution image reconstruction aims at obtaining a high-resolution image by fusing a set of low-resolution images. Low-resolution images are usually subject to some degradation which causes substantial information loss. Therefore, contiguous images in a sequence could be viewed as a degraded version (SR image) of an image at higher resolution (HR image). Using a multi-frame SR algorithm could achieve a restoration of the HR image. This work aims to investigate the possibility of using a multi-frame super-resolution algorithm to enhance the performance of a classic re-identification system by exploiting information provided by video sequences made available by a video surveillance system. In the case that the SR technique employed results in an effective performance enhancement, we intend to show empirically how many match frames are required to have an effective improvement.

1 Introduction

The creating, broadcasting and archiving of information in a video format is a growing phenomenon that is a direct consequence of the reduction in the cost of technology and the increase of the available network bandwidth. The availability of adsl/vdsl home connections with a large bandwidth has opened up a new class of services, such as IPTV, with the consequence that a large number of data streams need to be managed and organised. This is particularly true for video-surveillance systems, where a large amount of video data needs to be analysed, a requirement which has in the last few years stimulated research in the area of video analytics. In the field of video-surveillance a very important topic is

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 637–647, 2017.
https://doi.org/10.1007/978-3-319-68560-1_57

face recognition, both for the identification of the person in the scene, and for the re-identification of subjects from among different video footage captured at different times and/or places. The two problems have very different objectives. The former aims at assigning an identity to a subject detected in the scene. The target of the latter is to decide whether it is the same subject appearing in different videos, without considering the subject's identity. Despite the increase in camera resolution and video quality which has occurred over the last few years, the problem of image resolution in the context of face recognition is still an open topic of research. Many video surveillance systems do not use state-of-the-art technology and even the most modern systems are unable to recognize a face when the distance of the subject from the camera is great, with respect to the camera resolution. In this case, the face can occupy just a very small portion of the acquired image, with an effective resolution that is not sufficient for the recognition task.

Super-resolution methods can give a valid support to face-recognition systems that use low resolution video equipment, and can solve this problem fully, or partially. In the literature many different methods have been proposed [5,8,17], covering a wide range of activities. We can divide the different algorithms into two main macro-categories: single image methods, and multi-frame methods. The techniques falling into the first group try to increase the image resolution by using structures contained in the image itself; conversely, the techniques in the second group aim at obtaining an image with greater informative content by combining many observations of the same scene taken at different moments. A natural benchmark for all these algorithms is the classic technique of zooming (e.g., linear interpolation), which does not need any assumption on the image content. It is worth mentioning that a zoomed image obtained from an interpolation method has merely a larger number of pixels than the original, but this does not correspond necessarily to a greater informative content. However, it is the obtaining of better informative content that can be a key element for the improvement of the performance of a pattern recognition task, such as, in this case, face recognition. This paper introduces a super-resolution method in the pipeline of a biometric face-recognition task. In particular, it provides various innovative contributions compared to similar proposals already present in literature. First and foremost, two different super-resolution methods are compared: the first operating on a single image, the second based on the combination of consecutive frames. The second contribution is the analysis of the performance of the two super-resolution methods on two different face-recognition frameworks: the first is based on local feature extraction computed at pixel level, while the second works on patches of greater size with respect to a 3×3 mask.

The goal is to show how super-resolution techniques can have a better performance when using global recognition methods. Finally, the super-resolution method [14] adopted in this work does not require any face registration, which is a typical limitation of the majority of the techniques proposed in the literature.

The paper is structured as follows. Related works are briefly reviwed in the next section, Sect. 2. The architecture proposed is described in Sect. 3, and the experimental results are reported in Sect. 4. Section 5 concludes the paper with the final remarks.

2 Related Work

Over the last few years Person Re-identification has posed a significant challenge. One of the main difficulties is the low resolution of old cameras that can make every traditional technique to improve image quality unworkable. Person re-identification methods can be divided into two main groups:

- *single frame methods*, that aim to extract information about a person by analysing a single image;
- *multi-frame methods*, that use multiple images of the same person (usually obtained from one or more sequences) to build his/her signature.

For the first class, color and histogram-like methods have proved to be well suited for the retrieval of images with similar content, as in [6,8,11]; the main drawback of histogram based methods is the lack of any geometric or spatial information. In [5] the silhouettes of people are segmented into multiple horizontal stripes, and then color features are computed to characterize each segment. In [16] color features, together with a set of SURF points of interest, are extracted from the images and used to build a *person's descriptor*. Other techniques exploit the availability of other sources of information, such as the color of the clothes the subject is wearing, biometrics or collateral features, such as gait [15]. Such methods suffer from several drawbacks, like enlightenment sensitivity or pose changes, in addition to possible occlusions in the field of view.

On the other hand, multi-frame methods, like [10], collect several views from different cameras and build feature based on a variant of the SURF points. Authors in [1] adopt a cascade of grids of common region descriptors (e.g., SURF, SIFT). In [4] the use of a sequence of frames from a video instead of single still frames provides a significant increase in the performance.

In recent years the progress in camera technology, that can now record videos at a high resolution, has opened up the possibility of exploring unexploited paths in this research field, like the adoption of super-resolution algorithms in the face recognition [3,9,18] and re-identification pipeline. For the latter, previous works that try to take advantage of the SR algorithm include [2], that proposes a procedure for the recognition of low-resolution faces by using the features extracted from a high-resolution training employed as prior information in a super-resolution algorithm, and [13] that learns a pair of HR and LR dictionaries to generate a mapping function from the features of HR and LR training images. With the learned dictionary pair and mapping function, the features of LR images can be converted into discriminating HR features.

3 System Architecture

In a traditional video-surveillance system, different sequences taken from cameras are used to verify which identities declared for a set of people are true. Usually, the system is based on the individual biometric keys used to identify a single person; the key signatures are not invariant to sequence conditions and the image quality can be a factor that can affect the correctness of the identification. We aim to show how the introduction of a super-resolution algorithm in a classic face-based recognition framework can improve the performance. For this purpose, we add a super-resolution algorithm to a classic recognition pipeline (see Fig. 1), obtaining the following configuration:

1. two (or more) low-resolution cameras, each observing different not overlapping areas;
2. a super-resolution algorithm;
3. a bio-metric key producer (in our case a feature extractor);
4. a score function $S(k_1, k_2) \rightarrow \mathbb{R}$ used to indicate if the identification proposed is accepted or rejected by the system.

So, instead of using raw images taken from image sequences produced by cameras, we compute a higher quality image constructed by sequence fragments. More formally, given a gallery set G taken from image sequences of a set of people at a given resolution r_G, and a probe set P taken from image sequences at a given resolution r_P, our approach is based on extrapolating n contiguous frames from every sequence in G and generating a higher resolution image for each of them by using a given multi-frame super resolution algorithm. We indicate as G^n_{SR} the image set obtained. Subsequently, the same process is applied to P taking m contiguous frames from each sequence, and we indicate the resulting set as P^m_{SR}.

This process can be repeated varying m and n in order to compare performances. The resulting G_{SR} and P_{SR} sets are then used in a re-identification task with the two different matching schemes described in Sect. 3.2. The objective of this work is to show not only how SR can improve the performances of

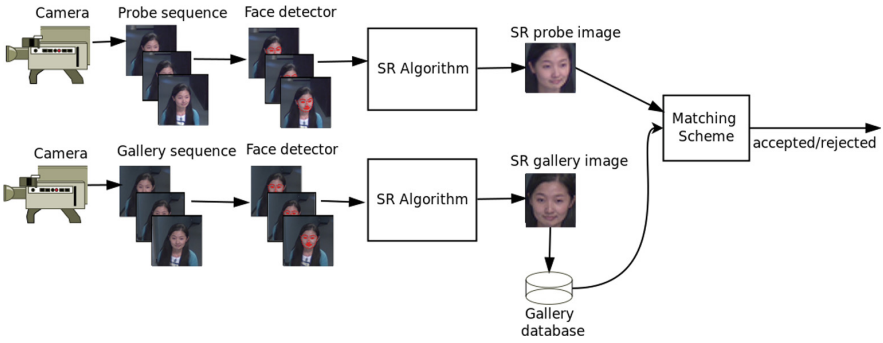

Fig. 1. Architecture of the system proposed.

a generic re-identification system, but also how the temporal information taken from different frames of the same sequence can affect the results.

3.1 Super-Resolution Algorithm

Our framework requires a multi-frame SR algorithm to synthesize the information provided by multiple frames of the same sequence in a single shot at a higher resolution. Super-resolution addresses the problem of reconstructing high-resolution data from a single or multiple low resolution observations. The key idea is based on the observation that every low-resolution image can contain different information on the same subject and that the fusion of these images can make it possible to extract subpixel information from the low-resolution image. Although the main focus of super-resolution methods is to obtain higher resolution images from low-resolution sequences, techniques of image restoration and image enhancement are also under consideration. SR techniques can be classified by using two parameters:

- methods that work in different domains (spatial/frequency)
- methods that work on the number of frames used in the restoration process (single/multiple)

Single frame methods can be considered equivalent to image interpolation, where the information in the image can be taken from the image itself. In our opinion, the state-of-the-art has been achieved by [12]; briefly, this study aims to search for similar patches in one image and computes the best homography which can generate the same image at a higher resolution. The following sequence of steps can clarify how it works:

Given an image I_O, to obtain an image I_H of a higher resolution of a k factor;

1. downsample the image of a k factor obtaining I_L image at a lower resolution
2. split images I_O and I_L in patches
3. for every patch P_O of I_O, compute a homography T s.t. $Q_L = T(P_O)$ where Q_L is the best patch matching I_L using a distance measure.
4. extract from I_O the Q_O patch that matches to Q_L in I_L
5. compute the inverse transformation T^{-1}
6. use T^{-1} to obtain P_H from Q_O, which is the patch in I_H that matches the position of P_O.

This first method is used to obtain SR single frame images.

Instead, multi-frame resolution enhancement is used when the information is taken from multiple frames. We can suppose that every image of the same subject owns inside it a certain information load; by contrast, multi.frame algorithms require a more complex pipeline to obtain significant results consisting of crucial intermediate steps like the registration between images; multi-frame methods often involve auxiliary algorithms taken from object detection or image registration. The authors in [14] propose a Bayesian method that seems to obtain good results, for which reason we have used it in our experiments. It aims to

estimate the best high resolution image that can generate a set of low-resolution images of the same scene using a probabilistic approach. In our experiments, we have compared the performances using a single-frame SR algorithm (i.e. [12]) and a multi-frame SR algorithm (i.e. [14]).

In Fig. 2 we show the output from both algorithms relative to a frame of a video sequence.

Fig. 2. Results of the SR algorithms considered. Top left to bottom right: orginal full resolution frame, result obtained using bi-linear interpolation, single frame SR, and multi-frame results using from 5 to 25 frames.

3.2 Feature Extraction and Matching Scheme

Once we have the SR images, we need a biometric key in order to discriminate in a unique manner a single face; our experiments have been performed with two different set-ups: the former using as the picture key-signature *LBP features* [2], a well-known type of global descriptor used in computer vision and face recognition, and the latter using the *spatial correlation index* [7].

– **LBP**: given an image I, the LBP operator assigns a label to every pixel by thresholding the 3×3-neighbourhood of each pixel with the central pixel value and considering the result as a binary number. Then, the labels histogram can be used as a descriptor. The facial image is divided into local regions and each descriptor is extracted from each region independently. All the descriptors are then concatenated to form a unique global face description $LBP(I)$.

  Once we have the keys for every SR_P and SR_G set, we need a matching scheme to compare the extracted bio-metric key; The similarity between two images I_1 and I_2 can be computed as the cosine similarity between $LBP(I_1)$ and $LBP(I_2)$.

– **Spatial Correlation**: the matching is performed by a localized version of the Spatial Correlation index. Given two images I_1, I_2 and the respective means $\mathbb{E}(I_1), \mathbb{E}(I_2)$, defining S as

$$s(I_1, I_2) = \frac{\sum_i \sum_j (I_1(i,j) - \mathbb{E}(I_1))(I_2(i,j) - \mathbb{E}(I_2))}{\sqrt{\sum_i \sum_j (I_1(i,j) - \mathbb{E}(I_1))^2 (I_2(i,j) - \mathbb{E}(I_2))^2}}$$

In our case, we divide all the images into subregions; for each subregion $r_1 \in I_1$ we search, in a subwindow around the same position in I_2, the region $r_2 \in I_2$ that maximizes $S(r_1, r_2)$. The global correlation is then obtained as

$$S_g = \sum_{r_1} s(r_1, (\arg\max_{r_2} s(r_1, r_2))).$$

The similarity between the two images I_1 and I_2 can be computed as $S_g(I_1, I_2)$.

4 Experimental Assessment

We decided to conduct an experimental evaluation of our proposed method on the publicly available ChokePoint video dataset[1]: this consists of videos of 29 subjects (23 male and 6 female) taken from different cameras. The videos in the dataset have a frame rate of 30 fps, and the frame resolution is 800×600 pixels. In total, the dataset consists of 48 video sequences, and 64,204 face images. In all the sequences, only one subject is present in every image at any one time.

4.1 Description of Experiments

The ChokePoint dataset gives us a set of contiguous frame sequences taken from a camera at a fixed resolution of 800×600; every sequence contains images of a moving person the face of whom can be enclosed in a bounding box of about 80×80; this resolution is too high for our purposes, so we resize every face to 40×40, obtaining what we consider our face-gold or Original Set $S_{original}$. Next, we randomly select a subset of 25 subjects and, for each of them, we take two distinct sequences, forming respectively the Gallery set $G_{original}$ and the Probe set $P_{original}$. We perform two sets of experiments, the former using the SR single-frame approach, the latter using the SR multi-frame approach. The performances are evaluated with the Genuine Acceptance Rate vs False Acceptance Rate (GAR-FAR) curves, and Cumulative Matching Curves (CMC). We subsample $G_{original}$ of a given factor d obtaining a set of low resolution sequences $G_{LR}^{\downarrow d}$. Next, we aim to restore the original resolution by applying an up-sample of the same factor using a standard interpolation algorithm obtaining the $G_{HR}^{\uparrow d}$ set. So, we can now use the SR algorithms to obtain our test-cases.

[1] Available from http://arma.sourceforge.net/chokepoint/.

The SR Single-Frame Approach: For every sequence in $G_{LR}^{\downarrow d}$, we select a frame where the face appears close to the camera and in a frontal pose, and choose this as a single frame. We then use the SR algorithm proposed by [12] to obtain a first set of super-resolved images G_{SR}^{1}. Analogously, we build P_{SR}^{1}.

The SR Multi-frame Approach: For every sequence in $G_{LR}^{\downarrow d}$, we select a sub-sequence (close to the conditions used for selecting the frame for the single-frame case) of the n_G contiguous frame that we want to combine together using the SR algorithm proposed by [14]; n_G is taken alternatively with values of $5, 9, 15$ and 20, obtaining $G_{SR}^{5}, G_{SR}^{9}, G_{SR}^{15}$ and G_{SR}^{20} sets. Analogously, we build $P_{SR}^{5}, P_{SR}^{9}, P_{SR}^{15}$ and P_{SR}^{20} sets.

Our objective is to evaluate how the identification performance changes when using

- SR images (P_{SR}, G_{SR});
- images at low resolution (P_{LR}, G_{LR});
- direct high-resolution images ($P_{original}$, $G_{original}$).

We use the distances discussed in Sect. 3.2, and we compare the G_{SR} sets versus the P_{SR} sets.

4.2 Results

As a first experiment we compared the performance when using the multi-frame SR with a different number of frames. Figures 3 and 4 show the results for the Spatial Correlation and LBP, respectively. The graphs show that using 5 frames gives, in general, the best performance. This is reasonable if we consider that the more frames we use, the more the pose of the subject can change, making the information *fusion* between the frames less coherent.

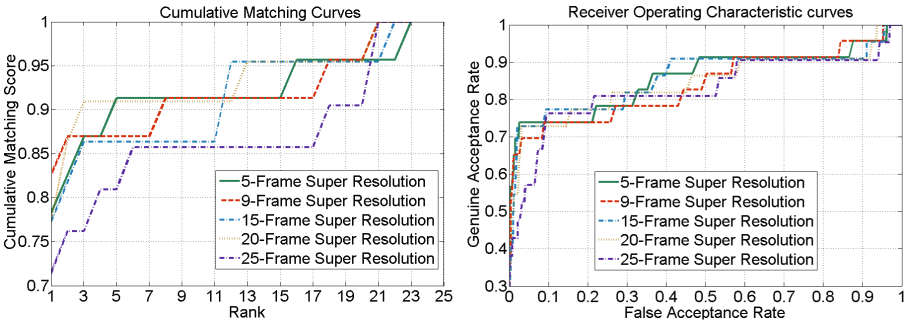

Fig. 3. Performance when using spatial correlation and the multi-frame SR algorithm with a different number of frames. Left: CMC curves. Right: GAR-FAR curves.

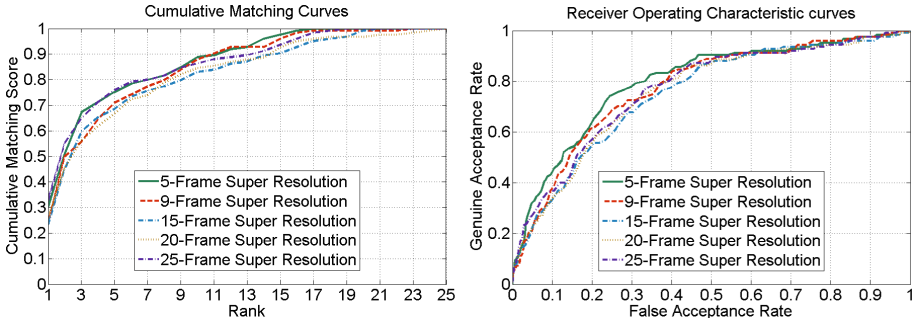

Fig. 4. Performance when using LBP and the multi-frame SR algorithm with a different number of frames. Left: CMC curves. Right: GAR-FAR curves.

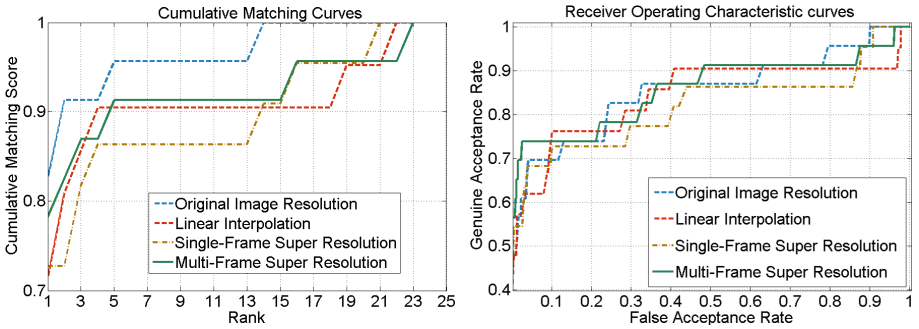

Fig. 5. Performance of the two SR algorithms against original HR images and interpolated images. Left: CMC curves. Right: GAR-FAR curves. Spatial correlation was used for this experiment.

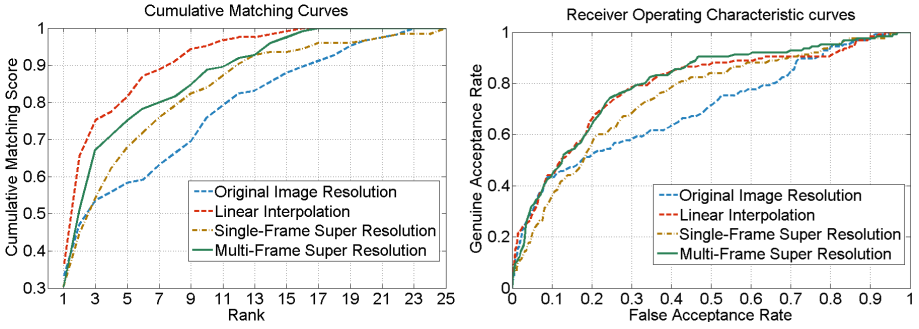

Fig. 6. Performance of the two SR algorithms against original HR images and interpolated images. Left: CMC curves. Right: GAR-FAR curves. LBP was used for this experiment.

In the next experiment we compared the performance of the multi-frame and single frame SR algorithms against the ones obtained from the original high resolution images. Moreover, in order to verify if the SR algorithm really does produce a positive result, we also measured the performance of the system when using images with a super-resolution with a simple interpolation procedure. In accordance with the results of the previous experiment, we used 5 frames for the multi-frame SR in this experiment. The results are shown in Figs. 5 and 6 for the Spatial Correlation and LBP, respectively. The results show a better performance when using spatial correlation. As for the resolution we can notice a slightly better performance when using the multi-frame SR than when the single-frame SR or the image interpolation is used. It can be also noted that, in general, the single-frame SR gives worse results than the image interpolation.

5 Conclusions

The work presented in this paper investigates the use of super-resolution in a face re-identification system. To achieve this objective we compared the performance of two different super-resolution methods, the former using only one frame for the super-resolution, the latter fusing the information from a sequence of frames. The performances have been measured using two different face-recognition frameworks: spatial correlation and LBP. The results obtained so far show that using a multi-frame super-resolution algorithm can give slightly better results. The experiments presented here also show a better performance when using spatial correlation. In the future our system needs to be tested on more difficult datasets than the one used for this study, and the experiments should be extended to a wider range of features.

References

1. Alahi, A., Vandergheynst, P., Bierlaire, M., Kunt, M.: Cascade of descriptors to detect and track objects across any network of cameras. Comput. Vis. Image Underst. **114**(6), 624–640 (2010). SpecialIssueonMulti-CameraandMulti-ModalSensorFusion. http://www.sciencedirect.com/science/article/pii/S1077314210000275
2. Baker, S.: Simultaneous super-resolution and feature extraction for recognition of low resolution faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, June 2008. https://www.microsoft.com/en-us/research/publication/simultaneous-super-resolution-and-feature-extraction-for-recognition-of-low-resolution-faces/
3. Bilgazyev, E., Efraty, B., Shah, S.K., Kakadiaris, I.A.: Improved face recognition using super-resolution. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. October 2011
4. Buml, M., Stiefelhagen, R.: Evaluation of local features for person re-identification in image sequences. In: 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 291–296. August 2011

5. Cong, D.N.T., Achard, C., Khoudour, L.: People re-identification by classification of silhouettes based on sparse representation. In: 2010 2nd International Conference on Image Processing Theory, Tools and Applications, pp. 60–65. July 2010
6. D'angelo, A., Dugelay, J.L.: People re-identification in camera networks based on probabilistic color histograms. In: 3DIP 2011 Electronic Imaging Conference on 3D Image Processing and Applications, January 2011, San Francisco, CA, USA, vol. 7882, pp. 23–27 (2011). http://www.eurecom.fr/publication/3274
7. De Marsico, M., Nappi, M., Riccio, D.: Face: face analysis for commercial entities. In: 2010 IEEE International Conference on Image Processing, pp. 1597–1600. September 2010
8. Forssen, P.E.: Maximally stable colour regions for recognition and matching. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. June 2007
9. Gunturk, B.K., Batur, A.U., Altunbasak, Y., Hayes, M.H., Mersereau, R.M.: Eigenface-domain super-resolution for face recognition. IEEE Trans. Image Process. **12**(5), 597–606 (2003)
10. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: 2008 Second ACM/IEEE International Conference on Distributed Smart Cameras. pp. 1–6. September 2008
11. Hatakeyama, Y., Makino, M., Mitsuta, A., Hirota, K.: Detection algorithm for color image by multiple surveillance camera under low illumination based-on fuzzy corresponding map. In: 2007 IEEE International Fuzzy Systems Conference, pp. 1–6. July 2007
12. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
13. Jing, X.Y., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., Hu, R., Xu, B.: Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2015
14. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 346–360 (2014)
15. Man, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 316–322 (2006)
16. de Oliveira, I.O., de Souza Pio, J.L.: People reidentification in a camera network. In: 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 461–466. December 2009
17. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Process. Mag. **20**(3), 21–36 (2003)
18. Wheeler, F.W., Liu, X., Tu, P.H.: Multi-frame super-resolution for face recognition. In: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–6. September 2007