

Deep Learning based Super-Resolution for Improved Action Recognition

Nasrollahi, Kamal; Guerrero, Sergio Escalera; Rasti, Pejman; Anbarjafari, Gholamreza; Baro, Xavier; J. Escalante, Hugo; Moeslund, Thomas B.

Published in:

International Conference on Image Processing Theory, Tools and Applications (IPTA), 2015

DOI (link to publication from Publisher):

[10.1109/IPTA.2015.7367098](https://doi.org/10.1109/IPTA.2015.7367098)

Publication date:

2015

Document Version

Early version, also known as pre-print

[Link to publication from Aalborg University](#)

Citation for published version (APA):

Nasrollahi, K., Guerrero, S. E., Rasti, P., Anbarjafari, G., Baro, X., J. Escalante, H., & Moeslund, T. B. (2015). Deep Learning based Super-Resolution for Improved Action Recognition. In *International Conference on Image Processing Theory, Tools and Applications (IPTA), 2015* (pp. 67 - 72). IEEE Signal Processing Society. <https://doi.org/10.1109/IPTA.2015.7367098>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Deep Learning based Super-Resolution for Improved Action Recognition

K. Nasrollahi*, S. Escalera[†], P. Rasti[‡], G. Anbarjafari[‡], X. Baro[§], H.J. Escalante[¶], and T.B. Moeslund *

* Visual Analysis of People laboratory, Aalborg University, Denmark e-mail: {kn, tbm}@create.aau.dk

[†] Human Pose Recovery and Behavior Analysis Group, University of Barcelona, Computer Vision Center, Spain
e-mail: sergio@maia.ub.es

[‡] iCv Group, Inst. of Technology, University of Tartu, Estonia e-mail: {pejman.rasti, shb}@ut.ee

[§] Universitat Oberta de Catalunya, Computer Vision Center, Spain e-mail: xbaro@uoc.edu

[¶] INAOE, Puebla, Mexico e-mail: hugojair@inaoep.mx

Abstract—Action recognition systems mostly work with videos of proper quality and resolution. Even most challenging benchmark databases for action recognition, hardly include videos of low-resolution from, e.g., surveillance cameras. In videos recorded by such cameras, due to the distance between people and cameras, people are pictured very small and hence challenge action recognition algorithms. Simple upsampling methods, like bicubic interpolation, cannot retrieve all the detailed information that can help the recognition. To deal with this problem, in this paper we combine results of bicubic interpolation with results of a state-of-the-art deep learning-based super-resolution algorithm, through an alpha-blending approach. The experimental results obtained on down-sampled version of a large subset of Hoolywood2 benchmark database show the importance of the proposed system in increasing the recognition rate of a state-of-the-art action recognition system for handling low-resolution videos.

I. INTRODUCTION

Action recognition is of great importance in many computer vision applications, such as human-computer interactions, automatic behaviour analysis, event detection, and abnormality detection. In real-world scenarios for computer vision applications like event detection or automatic behaviour analysis, video recordings of the scene of interest are usually provided by surveillance cameras. These cameras are usually installed in points which are few meters above the ground plane, to cover as much view of the scene as possible. This results in a large distance between the cameras and the objects of interest (here people). Consequently, people look very small in such surveillance videos as they occupy only few pixels in each frame of the videos. The larger the distance between the people and the cameras, the smaller the pictured people. This makes detection of the people and recognition of their actions extremely challenging. In order to deal with this problem, one could either use:

- manual observation and analysis of videos, being a time consuming and boring task.
- very high-resolution video recording devices, which are:
 - firstly, very expensive.
 - secondly, not that suitable for automatic video analysis as they generate a huge amount of data which are difficult to process and also to transmit over communication channels.

- use upscaling algorithms.

The problem with the simple upscaling algorithms, like bilinear interpolation or bicubic interpolation is that they produce artifacts which make their automatic analysis erroneous. Furthermore, they may not recover all the high-resolution details that can help the recognition process.

To produce upscaled high-resolution videos, that are less affected by artifacts, from low-resolution recordings, super-resolution algorithms have been used. These algorithms can be applied in both spatial and frequency domains [1]. The frequency domain approaches are more suitable for cases where motions of the objects follow simple models like translation. Examples of such cases are satellite imaging. For surveillance videos in real-world scenarios spatial domain approaches are more suitable as they can cope with the complicated motion structures of objects in the scene. These algorithms are generally divided into two groups: single-image-based and multiple-image-based super-resolution algorithms [1], [2].

Multiple-image super-resolution algorithms, like [3], receive a couple of low-resolution images of the same scene as input and usually employ a registration algorithm to find the transformation between them. This transformation information is then used along with the estimated blurring parameters of the input low-resolution images, to combine them into a higher scale framework to produce a super-resolved output image. For multiple-image super-resolution algorithms to work properly, there should be sub-pixels displacements between input low-resolution images. Furthermore, these sub-pixels displacements should be estimated properly by the registration algorithm, which is usually a challenging task, especially when complicated motion of non-rigid objects, like human body, needs to be modeled. These algorithms are guaranteed to produce proper higher resolution details, however, their improvement factors are usually limited by factors close to two [1].

Single-image super-resolution algorithms, like [4], do not have the possibility of utilizing sub-pixel displacements, because they only have a single input. Instead, they employ a kind of training step to learn the relationship between a set of high-resolution images and their low-resolution counterparts. This learned relationship is then used to predict the missing high-resolution details of the input low-resolution images. Depending on the relationship between the training low-and high-resolution images, these algorithms can produce high-










Down-sampling factor		
Original resolution	4	8
		
		
		
Recognition rates using bag of words (BoW)		
62.2%	30.73%	23.7%

Fig. 1: The importance of resolution in action recognition algorithms: recognition accuracy of the state-of-the-art action recognition algorithm of [9] drops as the resolution of images of different actions of the Hollywood2 dataset drops.

resolution images that are far better than their inputs, by improvement factors that are much larger than two [1].

Super-resolution algorithms have been applied to a wide range of applications, such as face recognition [5], [6], biometric recognition [7], and target recognition [8]. However, they have not been used for activity recognition in low-resolution videos. Such an application of super-resolution seems very important, because due to the distance between cameras and subjects of interest, in real-world scenarios, subjects occupy only a small portions of each frame, which makes activity recognition very challenging. Fig. 1 shows the importance and need for an upsampling algorithm in low-resolution images for action recognition algorithms. It is shown in this figure that, regardless of the targeted action, the recognition accuracy of action recognition algorithms drop considerably as the resolution of the images drop.

In this paper, we investigate the importance of the state-of-the-art deep learning-based Convolutional Neural Network (CNN) super-resolution algorithm of [10] in improving recognition accuracies of the state-of-the-art activity recognition algorithm of [9] for working with low-resolution images. To the extent of our knowledge, activity recognition in different resolutions has not been studied that much, except in Ahad et al. [11] in which appearance-based directional motion history image has been used to recognize various levels of video resolutions. However, our proposed system is the first one in which super-resolution algorithms have been employed for improving quality of low-resolution input images before action recognition. We show in this paper, that such super-resolution algorithms produce high-resolution details that are

not necessarily recovered by simple upscaling algorithms, like bicubic interpolation. It is shown in this paper, that combining the results of the deep learning-based super-resolution and the bicubic interpolation, through an alpha blending approach, produces images that are of better quality compared to the input low-resolution images. We show that employing such higher resolution images improves the recognition accuracy of a state-of-the-art action recognition algorithm.

The rest of the paper is organized as follows: related works are reviewed in Section 2. The proposed system and its sub-blocks are explained in section III. Experimental results are explained in section IV. Finally, section V concludes the paper.

II. RELATED WORKS

Super-resolution is playing an important role in many image processing applications such as satellite imaging [12], medical imaging [13], biometric recognition [14], and high dynamic range imaging [15]. Generally, super-resolution is employed either in frequency domain or spatial domain. In frequency domain approaches, the low-resolution image(s) are transformed to the frequency domain and then high-resolution estimation is obtained in that domain [16]. Wavelet transform is one of the most popular transformations which is used in frequency domain super resolution. This transformation is used to decompose the input image into structurally correlated subbands allowing exploiting the self-similarities between local neighbouring regions [1]. Existence of the-state-of-the-art techniques in super resolution has introduced a high impact into many image processing applications [17].

These days one of the most important research areas in human pose recognition is action recognition. In this field, the main focus is on realistic datasets collected from web videos, movies and TV shows [18], [19]. Feature trajectories and dense sampling play an important role in action recognition. Recent research [20] shows impressive results for action recognition by leveraging the motion information of trajectories. In [21], Harris3D interest points [22] were tracked with the KLT tracker [23] in order to extract feature trajectories. In their works an activity recognition feature was presented. The feature was based on velocity history of tracked key points. A generative mixture model was presented for video sequences and they illustrated that it performs comparably to local spatiotemporal features on the KTH activity recognition dataset. Sun et al. used SIFT descriptors to extract trajectories by matching them between two consecutive frames [24]. In their method the spatiotemporal context information was modelled in a hierarchical way. Efficient representations of intra-trajectory and inter-trajectory contexts were encoded into the transition matrix of a Markov process, and the extracted stationary distribution is used as final context descriptor. They recognize actions by employing the multichannel nonlinear SVMs. In [25], an approach was proposed to describe videos by dense trajectories. Dense points were sampled from each frame and tracked based on displacement information from a dense optical flow field. Their trajectories were robust to shot boundaries as well as fast irregular motions. The motion information in videos was covered by the dense trajectories. Authors also investigated how to design descriptors to encode the trajectory information. Finally, A novel descriptor robust

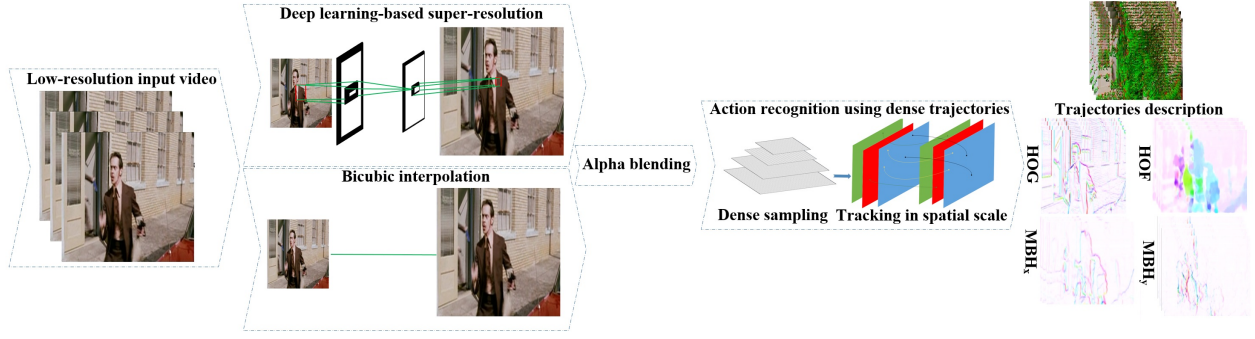


Fig. 2: The block diagram of the proposed system.

to camera motion based on motion boundary histograms was proposed. In [9], the authors presented improved dense trajectories model. In their model the camera motion was estimated by matching feature points between frames using SURF descriptors and dense optical flow. Finally, they remove this camera motion in order to improve their action recognition results.

III. THE PROPOSED SYSTEM

The block diagram of the proposed system is shown in Fig. 2. Having a low-resolution input video, the proposed system upsamples it by a bicubic interpolation and the deep learning super-resolution algorithm of [10] in parallel. Then, these two upsampled videos are combined using a simple alpha blending technique to produce a high-resolution video. Then, the dense trajectories of [9] are used to perform action recognition on the high-resolution video. These algorithms are briefly revisited in the following subsections.

A. Deep learning-based super-resolution applied to low-resolution videos

In this paper we use the super-resolution method of [10], which is based on a CNN with three layers. Given an image I , a low-resolution upsampled image \mathbf{Y} is created using bicubic interpolation. The first layer extracts overlapping patches from the image \mathbf{Y} and represents each patch as a high-dimensional vector. Instead of using pre-trained bases such as PCA, DCT or Haar, in this case these bases are optimized during the network optimization. This first layer can be expressed as an operation F_1 :

$$F_1(\mathbf{Y}) = \max(0, W_1 * \mathbf{Y} + B_1), \quad (1)$$

where W_1 and B_1 represent the filters and biases, respectively. In our case, W_1 is of a size $9 \times 9 \times 64$, where 9×9 is the spatial size of the filters and 64 is the number of the filters. B_1 is a 64 dimensional vector, where each element is associated with a filter.

The second layer performs a non-linear mapping of each of the 64-dimensional vectors from the first layer to a 32 dimensional vector. The operation performed by this second layer can be formulated as:

$$F_2(\mathbf{Y}) = \max(0, W_2 * F_1(\mathbf{Y}) + B_2), \quad (2)$$

where W_2 is of size $64 \times 5 \times 5 \times 32$ and B_2 is 32-dimensional. Each of those output 32-dimensional vectors is conceptually a representation of a high-resolution patch that will be used for reconstruction in the last layer. Although more convolutional layers can be added to increase the non-linearity, we use the configuration of [10], to avoid computational complexity of the method.

The output from the second layer is then passed to the last layer, where the reconstruction is performed. This layer emulates the classical method that averages the predicted overlapping high-resolution patches to create the final full image. This process can be formulated as:

$$F(\mathbf{Y}) = W_3 * F_2(\mathbf{Y}) + B_3, \quad (3)$$

where W_3 is of size $32 \times 5 \times 5$ and B_3 is a 1D vector (since we use single channel images). The logic after this formulation is that if the high-resolution patches are in the image domain, the filters act like an averaging filter, while if the representations are in some other domains, W_3 behaves like first projecting the coefficients onto the image domain and then averaging. In both cases, W_3 is a set of linear filters.

Learning the end-to-end mapping function F requires the estimation of parameters $\Theta = \{W_1, W_2, W_3, B_1, B_2, B_3\}$. Following the work of the same authors in [4], this is achieved through minimizing the loss between the reconstructed images $F(\mathbf{Y}; \Theta)$ and the corresponding ground truth high-resolution images \mathbf{X} . Given a set of high-resolution images $\{\mathbf{X}_i\}$ and their corresponding low-resolution images $\{\mathbf{Y}_i\}$, Mean Squared Error (MSE) is used as the loss function. This loss is minimized using stochastic gradient descent with the standard backpropagation. In particular, the weight matrices are updated as:

$$\Delta_{i+1} = 0.9 \cdot \Delta_i + \eta \cdot \frac{\partial L}{\partial W_i^l}, \quad W_{i+1}^l = W_i^l + \Delta_{i+1}, \quad (4)$$

where $l \in \{1, 2, 3\}$ and i are the indices of layers and iterations, η the learning rate (10^{-4} for $l \in \{1, 2\}$ and 10^{-5} for $l = 3$) and $\frac{\partial L}{\partial W_i^l}$ is the derivative. The filter weights of each layer are initialized by drawing randomly from a Gaussian distribution with zero mean and standard deviation 0.001 (and 0 for biases).









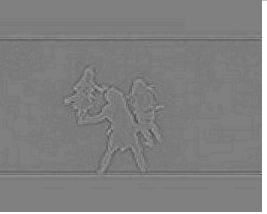




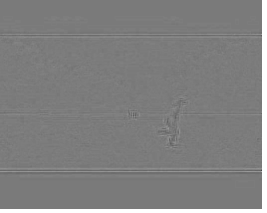











input	3x interpolation (bicubic)	3x super-resolved	difference	combined by $\alpha = 0.2$
				
				
				
				
				

Fig. 3: Few examples from Hollywood2 database (first column) and their upsampling counterparts by bicubic interpolation (second column) and deep learning-based super-resolution (third column). The difference between the two upsampling algorithms is shown in the fourth column. It can be seen from the difference images that the two upsampling algorithms do not necessarily produce similar high-resolution images, and hence complement each other (last column) for recognition purposes. This is further verified in the experimental results section.

B. Alpha blending of different upsampled videos

Following the block diagram of the proposed system shown in Fig. 2, having upsampled low-resolution videos by the bicubic algorithm and the above mentioned deep learning-based super-resolution algorithm, we need to combine them. To do so, we use the following alpha blending technique:

$$F_{HR} = \alpha * F_{SR} + (1 - \alpha) * F_{BI}, \quad (5)$$

in which, F_{SR} and F_{BI} are the high-resolution images produced by the super-resolution and bicubic algorithms, respectively. The reason for using these two different upsampling algorithms is that they produce different high-resolution details. This can be seen from the difference images (fourth column) of Fig. 3. These images have been obtained by subtracting the two upsampled images from each other, after mean filtering the super-resolved image by a kernel of size 3×3 . It can be seen from these difference images, that these two upsampling algorithms produce high-resolution images that are not necessarily the same. These high-resolution details mostly focus around the subject of interest, thus are important for action recognition. Having produced the high-resolution image, hereafter shown

by F , it is fed to the action recognition algorithm described in the next section.

C. Dense trajectories-based action recognition

This algorithm starts with a feature extraction step, in which the dense trajectories are considered as the criteria, which help determine the type of the action perceived. To do that, first, feature points are densely sampled on a grid spaced by K pixels, and then tracked in each spatial scale separately. A median filter is used to track each point $F_p = (x_p, y_p)$ from the p^{th} frame to the $p + 1^{\text{th}}$, in a dense optical flow field $W_t = (h_t, v_t)$, where v_t denotes the vertical component, and h_t stands for the horizontal component of the optical flow: $F_{p+1} = (x_{p+1}, y_{p+1}) = (x_p, y_p) + (M * W)|_{(x_p, y_p)}$, where M is the median filter kernel of size 33. It is worth mentioning that median filter is more robust compared to bilinear interpolation, and improves trajectories for points at motion boundaries. Afterward, the points associated with subsequent frames are concatenated to form trajectories in the form $(F_p, F_{p+1}, F_{p+2}, \dots)$, for each of which, five descriptors

are extracted: the trajectory itself, in other words, the concatenation of normalized displacement vectors, along with HOG, HOF, and MBH [9].

Dense trajectories and the corresponding descriptors are extracted from training and test videos, and the descriptors are normalized by means of the so-called RootSIFT approach, i.e. square root after L_1 normalization.

For the aim of this paper, following the same strategy as the one utilized in [9], 100,000 training descriptors were randomly selected for clustering, whereby, training and test videos were represented through the aforementioned bag-of-features representation. Then, feature spaces were combined by the classification model, which is tantamount to taking an SVM with a χ^2 kernel into account [9]. Finally, for the classification procedure, Fisher Vectors from the latter descriptors were concatenated so as to represent each video on the basis of a linear SVM.

IV. EXPERIMENTAL RESULTS

In order to show the effectiveness of the proposed system in improving the quality of low-resolution images, and hence increasing the recognition accuracy of the dense trajectories-based action recognition algorithm of [9], we performed the following two experiments on the Hollywood2 database. On the one hand, we assess the action recognition performance when processing low resolution videos. The goal of this experiment is to show how performance degrades when decreasing the resolution of videos. On the other hand, we evaluate the recognition performance after applying the super resolution method described in Section III. The goal is to assess the recognition rate improvement obtained when the quality of low resolution videos has been enhanced. In the following, we first explain the employed databases and then the details of the experimental results are given.

A. Data

The Hollywood2 database has been employed. The database includes 12 classes of human actions with 10 classes of scenes distributed over 3669 video clips. The actions are answer phone, drive car, eat, fight person, get out car, hand shake, hug person, kiss, run, sit down, sit up and stand up. The database contains various video clips from about 70 movies. The database is introducing a comprehensive benchmark for human action recognition in realistic and challenging settings and is used in experimental results in many state-of-the-art action recognition systems.

B. Details of the experimental results

Two experiments were performed for evaluation of the benefits of using the mentioned upsampling technique for improving the action recognition performance on low-quality videos. In the first experiment, we evaluated the recognition performance of the dense trajectories method using low-resolution videos of different resolutions. For this experiment, we down-sampled the employed databases by down-sampling factors of four and eight. Each of these down-sampling factors result in a new database in which we applied the action recognition method described in Section III-C. Results on this experiment on the entire database of Hollywood2 are shown

in Figure 1. It can be seen from this figure that the recognition rate generally drops as the resolution of the input images drops. This verifies the need for upsampling techniques for action recognition algorithms to work with low-resolution images.

In the second experiment we assess the benefits of using the alpha blending technique in combining the super-resolved videos obtained by the deep learning algorithm and those obtained by the bicubic interpolation. To do that, a subset of the Hollywood2 database has been chosen and down-sampled by a factor of two. This subset contains 53 videos for training and 59 videos for testing, covering all the actions of the database by at least eight to ten videos for each action of the database. The images in the down-sampled subset have been upsampled by a factor of two by both the bicubic interpolation and the deep learning-based super-resolution of [10], separately. The recognition rate of the dense trajectories-based super-resolution algorithm of [9] have been obtained for these two upsampling cases and are reported in Table I. Then, the images upsampled by these two upsampling techniques (bicubic interpolation and deep learning-based super-resolution) are combined using the mentioned alpha blending technique, with an (experimentally determined) alpha value of 0.2. Then, recognition rates of the dense trajectories-based super-resolution algorithm of [9] have been obtained for the images that are combined using the alpha blending technique. The results are shown in the last column of Table I.

The recognition results in Table I are reported for different descriptors obtained by the dense trajectories of [9], using a bag of words (BoW) technique. These descriptors are: the trajectory (TRT), HOG, HOF, MBH, and their combined version following the method of [9]. For the classification purpose, the code words for each of the descriptors are classified using a Neural Network (with one hidden layer of neurons, trained for 50 epochs, with learning rate of 0.1) and a linear SVMs. It can be seen from this table that the proposed upsampling techniques using the alpha blending produces better results compared to the other two cases using the Neural Network classifier. This verifies that deep learning-based super-resolution algorithms and other upsampling techniques (like bicubic interpolation) can complement each other in enhancing the quality of videos and hence improve the recognition performance of the state-of-the-art action of [9]. It should be also noted that the best performance in this paper has been obtained using the Neural Network with the TRT, while in the work of [9] the best recognition rate has been obtained using SVM from the combined classifier. However, the system of [9] uses the original images, while we have first down-sampled and then upsampled them. This means that the proposed combined classifier of [9] is not robust against down-sampling compared to TRT.

V. CONCLUSION

State-of-the-art action recognition algorithms, like dense trajectories of [9] have difficulties handling videos that are of low quality and resolution. Such videos are however very common in real-world scenarios, for example, in videos captured by surveillance cameras. To deal with such cases in this paper we have proposed to use upsampling techniques. We have found that the state-of-the-art deep learning-based super-resolution algorithm of [10] and bicubic interpolation

TABLE I: The recognition accuracy of the dense trajectories-based action recognition algorithm of [9] for a subset of images of the Hollywood2 database that are first down-sampled by a factor of two and then are upsampled by different upsampling algorithms, including the one proposed in this paper.

Classification method	Descriptor	Upsampling method		
		Interpolated by bicubic	Super-resolved by deep learning	combined by alpha blending
Neural Network	TRT	36.78 %	36.18%	39.28%
Neural Network	HOG	24.92%	24.02%	26.98%
Neural Network	HOF	23.62%	23.16%	20.22%
Neural Network	MBHx	20.37%	21.76%	25.69%
Neural Network	MBHy	20.51%	20.45%	19.80%
Neural Network	combined	29.87%	23.57%	34.47%
Liner SVM	TRT	13.74%	11.26%	13.10%
Liner SVM	HOG	11.47%	14.18%	13.32%
Liner SVM	HOF	14.19%	11.97%	11.32%
Liner SVM	MBHx	12.78%	13.06%	11.21%
Liner SVM	MBHy	10.33%	12.28%	12.92%
Liner SVM	combined	14.04%	13.73%	14.50%

complement each other and each of them produces high-resolution details that are not necessarily produced by the other one. Hence, we combine the results of these two through an alpha blending. Experimental results on a down-sampled version of the Hollywood2 benchmark database show that the proposed is efficient in improving the quality of such low-resolution videos and hence improves the recognition accuracy of the dense trajectories action recognition algorithm of [9].

REFERENCES

- [1] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.
- [2] P. Rasti, H. Demirel, and G. Anbarjafari, "Image resolution enhancement by using interpolation followed by iterative back projection," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*. IEEE, 2013, pp. 1–4.
- [3] G. Polatkan, M. Zhou, L. Carin, D. M. Blei, and I. Daubechies, "A bayesian nonparametric approach to image super-resolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, pp. 346–358, 2015.
- [4] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 184–199.
- [5] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *Neural Networks, IEEE Transactions on*, vol. 22, no. 1, pp. 121–130, 2011.
- [6] K. Nasrollahi and T. Moeslund, "Finding and improving the key-frames of long video sequences for face recognition," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, 2010, pp. 1–6.
- [7] J. Cui, Y. Wang, J. Huang, T. Tan, and Z. Sun, "An iris image synthesis method based on pca and super-resolution," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4. IEEE, 2004, pp. 471–474.
- [8] K. Feng, T. Zhou, J. Cui, and J. Tan, "An example image super-resolution algorithm based on modified k-means with hybrid particle swarm optimization," in *SPIE/COS Photonics Asia*. International Society for Optics and Photonics, 2014, pp. 92 731I–92 731I.
- [9] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3551–3558.
- [10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015. [Online]. Available: <http://arxiv.org/abs/1501.00092>
- [11] M. A. Ahad, J. Tan, H. Kim, and S. Ishikawa, "A simple approach for low-resolution activity recognition," *Int. J. Comput. Vis. Biomech*, vol. 3, no. 1, 2010.
- [12] H. Demirel and G. Anbarjafari, "Discrete wavelet transform-based satellite image resolution enhancement," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 6, pp. 1997–2004, 2011.
- [13] H. Greenspan, "Super-resolution in medical imaging," *The Computer Journal*, vol. 52, no. 1, pp. 43–63, 2009.
- [14] G. Fahmy, "Super-resolution construction of iris images from a visual low resolution face video," in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*. IEEE, 2007, pp. 1–4.
- [15] T. Bengtsson, I.-H. Gu, M. Viberg, and K. Lindstrom, "Regularized optimization for joint super-resolution and high dynamic range image reconstruction in a perceptually uniform domain," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1097–1100.
- [16] N. Bose, H. Kim, and B. Zhou, "Performance analysis of the tls algorithm for image reconstruction from a sequence of undersampled noisy and blurred frames," in *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, vol. 3. IEEE, 1994, pp. 571–574.
- [17] N. Othman, N. Houmani, and B. Dorizzi, "Quality-based super resolution for degraded iris recognition," in *Pattern Recognition Applications and Methods*. Springer, 2015, pp. 285–300.
- [18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
- [19] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 12, pp. 2441–2453, 2012.
- [20] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 514–521.
- [21] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 104–111.
- [22] I. Laptev and T. Lindeberg, "Space-time interest points," in *Computer Vision, 2003 IEEE 12th International Conference on*. IEEE, 2003, pp. 432–439.
- [23] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, vol. 81, 1981, pp. 674–679.
- [24] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2004–2011.
- [25] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.