# Convolutional Sparse Coding of Images through Characteristic Points Extraction

Damian Salata, Arkadiusz Tomczyk[0000−0001−9840−6209], and Piotr S. Szczepaniak[0000−0002−9973−0673]

Institute of Information Technology
Lodz University of Technology
al. Politechniki 8, 93-590 Lodz, Poland
arkadiusz.tomczyk@p.lodz.pl

**Abstract.** Sparse coding refers to methods that allow to reveal, in an unsupervised way, internal structure hidden in the data. When applied to images, it enables their automatic decomposition into a set of semantic elements. There are several classic algorithms that allow to address sparse coding problem. In this work we show that one can adapt for that purpose a convolutional neural network, which originally was used to solve completely different task (characteristic points extraction). Our contribution includes both network architecture and a method of its training.

**Keywords:** Sparse coding · Characteristic points · Convolutional neural networks · Image analysis.

## 1 Introduction

Sparse coding and characteristic points (landmarks) extraction are two, usually not connected, groups of techniques which find multiple applications in image analysis. The first ones are an unsupervised learning algorithms able to discover a set of basis functions that capture higher-level features in the data. When image data are taken into account, images are automatically decomposed into a set basic elements, which allows to reveal their internal structure and consequently leads to their alternative, reduced representation. It finds its application in: image classification, noise removal, compression, image synthesis, etc. Moreover making sparse codes is considered as a plausible model of the visual cortex ([4]). The second, are methods for detecting and describing local features of the images. Well known classic techniques used for that purpose are: SIFT, SURF, ORB or BRISK ([6, 1, 7, 5]). Characteristic points were initially based on corners, which are relatively easy to find. In recent years, however, development of convolutional neural networks CNN ([3]) made it possible to use them for this task as well. This allowed to extract characteristic points, which did not necessarily have to be corners. One of the first algorithms based on CNNs, which simultaneously determined characteristic points and their descriptors, was SuperPoint ([2]). Characteristic points are used for: object detection, 3D reconstruction,

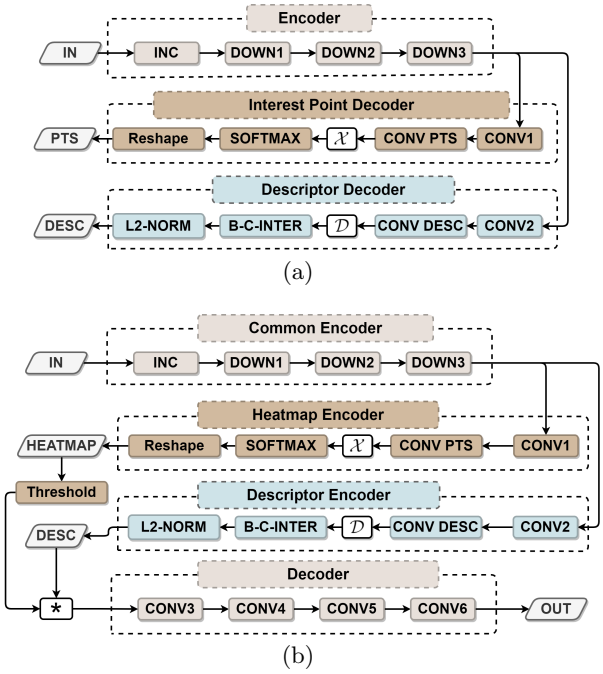


**Fig. 1.** The difference between original SuperPoint architecture (a) and its modified version used in this work (b). The main difference is abandonment of characteristic point extraction PTS and combination of thresholded HEATMAP with DESC to create the input of the decoder (they contain sparse code of the input image).

panorama stitching by combining different images, object tracking, video stabilization, etc.

In this paper we propose a novel method for image sparse coding which bases on characteristic points extraction algorithm. The original contribution of this paper covers: encoder basing on landmark detector, 2D sparse codes with non-linear decoder, loss functions and two-stage training procedure.

## 2 Method

In our approach we use a modified SuperPoint[1] convolutional characteristic points extractor, which architecture is presented in Figure 1a, as an encoder part of autoencodoer shown in Figure 1b. This architecture was chosen since its working principle enables an easy imposition of sparsity constraints. The original method for a given input image of size $W \times H$ generates two outputs $\mathcal{X}$

[1] Due to limited number of pages, only necessary details of this method are presented in this work - we encourage reader to get familiar with original work [2].

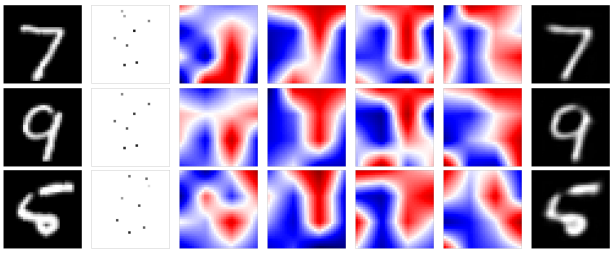

**Fig. 2.** Sample reconstructions of selected MNIST test images, from left there are: input image, thresholded HEATMAP, all channels of DESC map and network output.

and $\mathcal{D}$ of reduced size $W/8 \times H/8$, which after further processing, lead to positions of landmarks PTS and $d$-dimensional descriptor map DESC (after bicubic interpolation size of DESC is the same as the size of input image).

Our modifications were motivated by both the need of adding convolutional decoder and the necessity of proper loss function $\mathcal{L}$ definition. We do not extract PTS from $\mathcal{X}$ but generate HEATMAP containing probabilities of landmarks existance separately in non-overlapping $8 \times 8$ blocks. In order to prevent decoder learning from insignificant codes we threshold the HEATMAP to remove small values (values below 0.1 are set to be 0). Finally, since both thresholded HEATMAP and DESC are of the same size $W \times H$, we multiply them elementwise to create the input of the decoder. To define sparsity constraints we use properties of $\mathcal{X}$ ([2]). It contains vectors with 65 elements. First 64 values are later interpreted as probabilities of landmarks in HEATMAP blocks (after reshaping), whereas the last one indicates the probability of landmark absence in the considered block. Since, within every block, the sum of all those probabilities is equal to 1 (thanks to the softmax operation) to force minimum number of landmarks (sparsity) we can expect a value close to 1 at the last position of $\mathcal{X}$ vectors. To express less restrictive requirement of having only one or none significant landmark, we can anticipate that maximum value of vectors in $\mathcal{X}$ should be close to 1. Further, we denote constraints taking into account the above observations as $\mathcal{L}^{hard}$ and $\mathcal{L}^{soft}$, respectively.

Autoencodoer is trained to reconstruct its input on the output. That is why the main component of the loss function is a mean squared error $\mathcal{L}^{mse}$ between network output and input. The preliminary experiments have shown that direct usage of hard sparsity constraint does not allow finding the desired optimum. That is why, we propose a two-stage strategy where at the begining network is trained using $\mathcal{L} = \mathcal{L}^{mse} + \lambda \cdot \mathcal{L}^{soft}$, and later the $\mathcal{L} = \mathcal{L}^{mse} + \lambda \cdot \mathcal{L}^{hard}$ is applied.

## 3 Results

To show the properties of the proposed approach we have trained autoencoder using MNIST dataset with hand-written digits. In the conducted experiments the descriptor size $d$ was equal to 4 and $\lambda = 0.1$. Sample reconstruction results

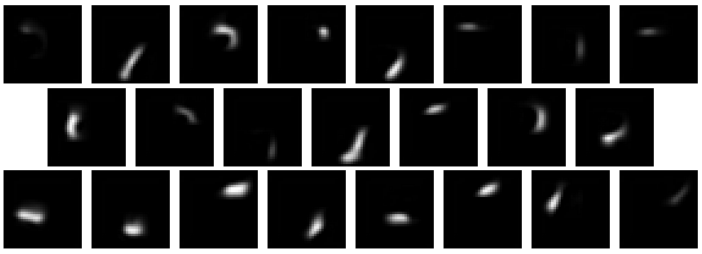

**Fig. 3.** Visual interpretation of discovered codes for images presented in Figure 2 (landmark distribution as well as their number are different for different images).

together with codes (thresholded HEATMAP) and descriptor maps (DESC) are presented in Figure 2. To show how images are decomposed or, in other words, to understand what is a visual interpretation of automatically discovered codes, we have used a decoder to process every landmark separately (Figure 3).

## 4 Summary

In this work we have shown how modified architecture of SuperPoint network, originally used for characteristic point extraction, can be applied for sparse coding of the given class of images. The proposed method was tested using relatively simple MNIST dataset. Further experiments will focus on more complex datasets and on the usage of generated, reduced image representations for image analysis related tasks such as: classification, segmentation, etc. In particular we are interested in application of graph convolutional neural networks able to directly operate on alternative representations of image content.

## References

1. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. vol. 110, pp. 404–417 (01 2006)
2. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. CoRR **abs/1712.07629** (2017)
3. Lecun, Y., Bengio, Y.: Convolutional Networks for Images, Speech and Time Series, pp. 255–258. The MIT Press (1995)
4. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems. vol. 19. MIT Press (2006)
5. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. In: 2011 International Conference on Computer Vision. pp. 2548–2555 (2011). https://doi.org/10.1109/ICCV.2011.6126542
6. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 1150–1157 vol.2 (1999). https://doi.org/10.1109/ICCV.1999.790410
7. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an efficient alternative to sift or surf. pp. 2564–2571 (11 2011). https://doi.org/10.1109/ICCV.2011.6126544