INVARIANT REPRESENTATION FOR BLUR AND DOWN-SAMPLING TRANSFORMATIONS

Huxiang Gu[⋆†] *Leibo Joel*[†] *Anselmi Fabio*[†] *Chunhong Pan*[⋆] *Tomaso Poggio*[†]

[⋆]Institute of Automation Chinese Academy of Science
[†]Massachusetts Institute of Technology

ABSTRACT

Invariant representations of images can significantly reduce the sample complexity of a classifier performing object identification or categorization as shown in a recent analysis of invariant representations for object recognition. In the case of geometric transformations of images the theory [1] shows how invariant signatures can be learned in a biologically plausible way from unsupervised observations of the transformations of a set of randomly chosen template images. Here we extend the theory to non-geometric transformations such as blur and down-sampling. The proposed algorithm achieve an invariant representation via two simple biologically-plausible steps: 1. compute normalized dot products of the input with the stored transformations of each template, and 2. for each template compute the statistics of the resulting set of values such as the histogram or moments. The performance of our system on challenging blurred and low resolution face matching tasks exceeds the previous state-of-the-art by a large margin which grows with increasing image corruption.

Index Terms— invariant representation, non-geometric transformations, blur, down-sampling

1. INTRODUCTION

Learning transformation invariant representations is now thought to be the crux of object recognition and thus of importance not just for computer vision, but also for neuroscience [1, 2]. Despite recent high profile successes in object and face recognition, there are several situations in which existing approaches are not robust enough compared to human vision. An interesting example is represented by images that are blurred due to either an out-of-focus lens, or atmospheric turbulence, or relative motion between the sensor and targets. Similar difficulties arise with low resolution images taken by surveillance cameras, or from a large distance. In face recognition there is also the additional problem that the gallery images to be matched with are typically high resolution, e.g., passport photos, but the probe images are poor quality surveillance camera frames [3, 4]. The problem can be cast as one of synthesizing an image representation that is invariant to these quality-decreasing transformations. We take this perspective here.

Alternative approaches address this problem by constructing a de-blurred or high resolution image using de-blur [5, 6, 7] or super resolution methods [8]. These methods solve a harder problem than the one that is strictly needed here, trying first to obtain a perceptually good reconstruction, and then to use it for recognition. Arguments based on the accumulation of errors through multi-step processes, and/or the data processing inequality suggest that these methods may not be optimally robust. Moreover, most of these methods take advantage of prior information from the data such as the fact that it is a face. Thus these methods are class specific [9, 10, 8, 11].

A result in Poggio et al [1] (see also [12]) theory of transformation invariant hierarchical architectures is that for transformations that form a locally compact group, such as translation, scale [13] and rotation in plane [14], it is always possible to obtain an invariant signature in a generic way, that is, class-specific templates are not required. We extend these results to non-geometric transformation by showing that several kinds of blur and down-sampling transformations satisfy the key hypotheses of the theory. We illustrate that this construction is possible whenever the transformation can be expressed as a convolution with a linear, symmetric[1] filter (we call it Linear Symmetric Convolution or short for LSC transformation below).

Our method can also extract invariant features robust to blur or down-sampling resolution [15, 5, 16]. The current state of the art for these problems is either class-specific or supervised: for example, Soma et al [15]'s system simultaneously maps the poor quality probe images and high quality gallery images into a subspace in which distance between images from the same subject is smaller than the original space. Her model is a supervised model which means that face data is needed for training and limited for face application. Moreover, her method is trained differently for different probe resolutions which dramatically affect its application since the actual resolution in real applications is unknown. Gopalan et al

[1]Here we mean center symmetric, the upper left corner equals the bottom right corner, such as [0 0.25 0; 0.25, 0.5, 0.125; 0 0.125 0]. This requirement guarantees (see the proof of 1 in Appendix) that the corresponding operator is self-adjoint.

[16] proposed a blur descriptor which is robust to both homogeneous and spatially varying blur. Unfortunately, one of his three assumptions that there is no noise in the system is not practical. Above all, neither of these algorithms address the problem of blur and down-sampling transformations jointly, nor do they extend to affine and illumination transformation easily.

We highlight our contributions in the following three aspects. First, we theoretically and experimentally prove that invariant signature of the LSC transformation can be achieved: two typical examples of LSC transformation are illustrated. We can compute invariant representations robust to blur and down-sampling transformations in the same framework. Second, we show empirical demonstrations of the theoretical result that LSC transformations are generic. That is, we can use templates from one class (even noise patches!) and obtain a representation which is invariant even when tested on other classes. Lastly, although our model only involves two simple steps – inner product and pooling – we also propose a method to speed up the computation (without sacrificing performance) by skipping the computation of dot products that turn out not to contribute to the final solution.

2. THEORY

M-theory [1] – as well as one of its specific and partial implementations (the HMAX model [3]) is based on a cascade of Hubel-Wiesel (HW) modules [17] which consists of two layers: simple cells and complex cells. The simple cell (S-unit) implements a dot product with a stored transformation of a template t, while the complex cell (C-unit) implements the pooling operation. We now explain how by these two operations we can compute a transformation-invariant signature under certain conditions for the case of the blur transformation.

Consider an image blurred by convolution with a Gaussian filter with variance σ^2 which for simplicity here is supposed to be discretized between 1 and N ($\sigma^2 = j$). We show that pooling with the max operation using a set of templates blurred to different degrees provides a signature that has partial blur invariance, that is invariance for certain ranges of blur of the input image. Our algorithm is based on the following proposition

Proposition 1. *Let $I, t \in L^2(\mathbb{R})$. Let $t^k = g_k * t$ with $g_k(x) = e^{-\frac{x^2}{k}}$, $k = 1, ..., N$. Let $u(I) = max_{k\in[0,\cdots n]} \langle I, t^k \rangle$. Further let $k^* = argmax_{k\in[0,\cdots n]} \langle I, t^k \rangle$. We have*

$$u(I) = u(I^j), \ \ \forall\, j < k^*, \ \ I \in L^2(\mathbb{R})$$

The proof is simple (see supplementary material). The result shows that we only need to store a group of templates t^k with $k = 1, \cdots, K$ and all of its blurred version (over a range of blur variances) instead of observing and storing all of the possible blurred version of I. This follows the paradigm of the M-theory approach in which given a novel image, we can use the full group of several templates $t^1, t^2, \ldots, t^K$ and their transformations (under G) to compute a signature $u(I)$ which is invariant to $g \in G$. Previous work developed the theory for geometric transformations and applied it to the affine group in the plane (translation, scale and rotation [23, 19-20]) and to non group smooth, class-specific transformations (such as change of pose of a face or a body). The work here is the first to apply a similar approach to non-geometric transformations. The proposition above suggests the algorithmic pipeline shown in Figure 1. Note that the templates are independent from the input image (the transformation is generic and not class-specific) so that the templates can be almost any image including noise patches.

3. EXPERIMENTS

In this section, we show a set of experiments on Extended Yale Face Database B [18] and Multi-PIE database [19]. Figures are better viewed when magnified.

In the first set of experiments, we test two LSC transformations, namely Gaussian-blur and down-sampling, on the extended Yale Face Database B. This dataset contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. Since pose and illumination are class specific transformation, we only select 8 frontal faces per human subjects here to focus on LSC transformation. There are only some subtle illumination or expression changes between probe images and gallery images, as shown in Figure 2 Panel II. The Gaussian kernels is of size 31×31 with ranging from 1 to 21 in steps of 2 (11 variations), so the pooling range of each templates is 88. While the resolution of input image for down-sampling test varies from 96×84 to 6×4 (9 variations), so the pooling range of each templates is 72.

Using the max pooling procedure, we find, as expected, that the max value always comes from the same row, as shown in Figure 1. In the dataset of Figure 2 we experimentally find that 90.1 percent of the max value is from the same row and 99.1 percent is from the same or adjacent rows. Therefore we can first compute the inner product between the input image and the full orbit of one template and find the max value (for instance, the j-th inner product); then we only need compute the inner product between the input image and j-th image of the full orbit of other templates. In this way, we can skip the computation of most dot product and pooling operation. Roughly speaking, if the number of templates is K and each full orbit has M images, then this simplified algorithm reduces the complexity of our model from $K \times M$ to M. From the comparison between the cyan curve and red curve of Figure 2, we can see that this simplified model achieves comparable recognition accuracy with the original model.

Interestingly, high recognition accuracy (95%) can still be achieved even with face images that are down-sampled to

Fig. 1. Pipeline of the proposed algorithm. There are two steps. First, the normalized dot product between the input image and each one of the templates of the i-th subject (thought to be stored in memory) is computed. Second, the max value (pooling function in all the following experiments is max pooling.) over the set of these inner products gives the i-th component of the representation of the input image. These two steps are repeated for each template until a K-dimensional representation is produced.

6×5, as shown in Figure 2 Panel I (Blur curve in left chart) in which noisy patches are adopted as templates. Further experiments on varying amount of noise patch templates can be found in Figure 3 A.

In the second set of experiments, we first analysed the performance of the proposed model with varying amount of templates on subset of YaleB dataset. Since we can train our model even on noise patch images, we can create as many templates as possible. From Figure 3 A, we can find that we can achieve perfect invariance across different standard deviations while preserve the identity information using 100 templates.

Then we compared our model with the state-of-the-art on blur invariance. The experimental set-up is similar with Taheri[2] [20]. We resized all images to 64×64, added 30dB white Gaussian noise to test robustness and created eleven different synthetically blurred sets of images using Gaussian kernels. However, we test our model on more severe cases: we adopt kernels of size 3131 with ranging from 1 to 21 in steps of 2 into our templates. In order to be more practical, the standard deviations of the input image are from 2 to 20 in steps of 2. The recognition accuracy are presented in Figure 3 B, where we compare with Taheri's model [20] and two existing deblurring-based methods [7, 6]. We can see that our model achieve perfect invariance even under severe conditions without knowing the Gaussian kernels before. Due to limited pages, we show more results on motion blur in our supplementary material.

In the last part, we further compare our model with Soma's algorithm across down-sampling transformation and illumination variation. This experiment is performed on CMU Multi-PIE face dataset [19] which contains images of 337 subjects. All subjects were taken under 15 view points and 20 illumination conditions while displaying a range of facial expressions. We randomly choose 200 subjects which have neutral expression and 20 illuminations for training and the rest for testing. The final recognition accuracy is achieved by averaging 20 different illumination conditions. The resolution of gallery image is 48×40 while probe image varies from 48×40 to 6×5. There are 6 resolution variation and 20 illumination conditions, so pooling range in this experiment is 120. Figure 4 shows the recognition accuracy on different probe image resolution. We can see that our model can get a good result even when probe images are down-sampled to 6×5. Compared with Soma's algorithm, our model can achieve much better result on very low resolution images and the margin grows with increasing image corruption. Despite the better performance, our model is neither class-specific nor resolution-specific.

[2]Note that our result is based on YaleB dataset instead of FERET dataset. The probe image has some subtle illumination and expression variation compared to the gallery image.

4. CONCLUSION

In this paper, we illustrate that partially invariant signature can be obtained also in the case of non-geometric transformations, in particular blur transformations, which are not a group. Despite its extreme simplicity, our algorithm shows impressive performance on LSC transformation even when the unsupervised learning only uses noisy patches.

5. REFERENCES

[1] T. Poggio, J. Mutch, F. Anselmi, J. Z. Leibo, L. Rosasco, and A. Tacchetti, "The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work)," 2012.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Lake Tahoe, CA, 2012, vol. 25.

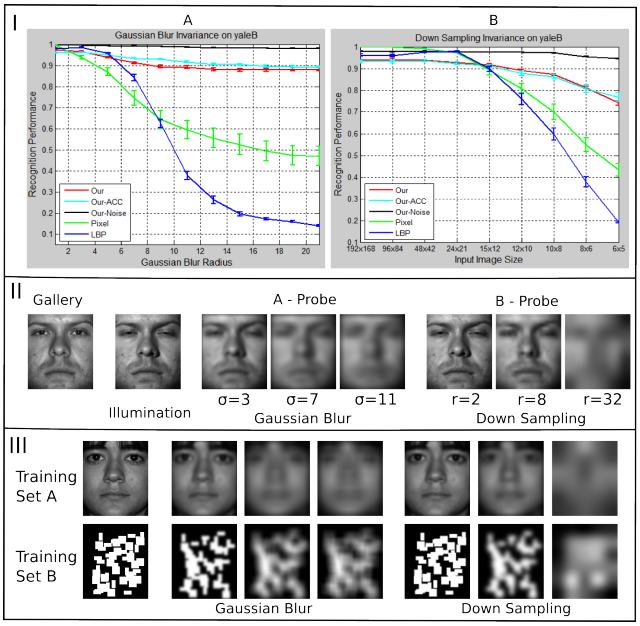

Fig. 2. Gaussian-blur (left) and down-sampling (right) invariance on a subset of Extended Yale Face Database B [18]. The target is to match the probe faces to the gallery faces. For each human subject, we select eight images which have some subtle illumination and expression changes. For Gaussian-blur, all images are convoluted with the Gaussian kernels of size 31×31 and radius ranging from 1 to 21 in steps of 2. For down-sampling, we resize the input image from 192×168 to 6×5 using standard bicubic interpolation. All down-sampled images are up-sampled to the same size of the gallery images with the same interpolation. Panel I shows recognition performance while panel II and III illustrate testing and training examples respectively. The red curve in panel I shows the recognition performance on training set face (randomly choose 18 human subjects for training and the rest 20 for testing) while the black curve on training set noise (randomly choose 30 noise images for training and 20 human subjects for testing). The cyan curve shows the performance of our accelerate model and the black curve denotes performance of LBP descriptor [20]

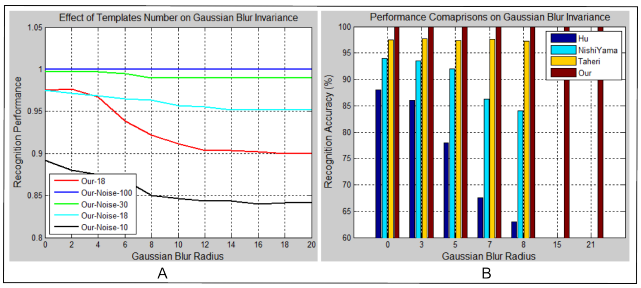

Fig. 3. Further experiments to illustrate Gaussian-blur invariance on Extended Yale Face Database B [18]. Left chart (A) shows performance on different number of templates. The red curve shows the performance of training on 18 faces and curves of other color illustrate the performance of training on noisy patches with templates number varying from 100 to 10. We can see that better recognition results can be achieved by using noise patches which are infinite rather than face data. Right chart (B) compares our model with algorithm from Hu [7], Nishiyama [6], and Taheri [20]. We can see that our model achieve better recognition performance even under severe conditions without knowing the Gaussian kernels before.

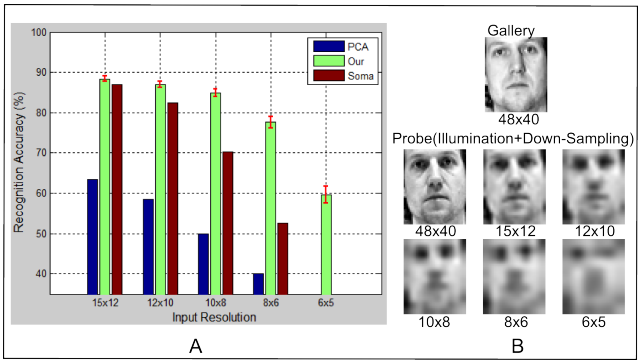

Fig. 4. Recognition performance comparisons with Soma's model [15] (A) under different probe resolution (B). Our model can achieve much better result on very low resolution images (such as 6×5) and the margin grows with increasing image corruption.

[3] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," 2003, vol. 35, p. 399458.

[4] P. Hennings-Yeomans, S. Baker, and B. Kuma, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," *IEEE Conf. on Computer Vision and Pattern Recognition*, p. 18, 2008.

[5] V. Ojansivu and J. Heikkila, "A method for blur and affine invariant object recognition using phase-only bispectrum," 2008, p. 527536.

[6] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference using subspace analysis for recognition of blurred faces," 2011, vol. 33, p. 838845.

[7] H. Hu and G. de Haan, "Low cost robust blur estimator," 2006, p. 617620.

[8] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," 2002, vol. 24, p. 11671183.

[9] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation.," 2009.

[10] A. Chakrabarti, A.N. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel pca-based prior," 2007, vol. 9, pp. 888–892.

[11] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," 2008, pp. 1–4.

[12] Anselmi F, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," 2013.

[13] J. Z. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio, "Learning generic invariances in object recognition: Translation and scale," 2010.

[14] J. Z. Leibo, J. Mutch, and T. Poggio, "Why the brain separates face recognition from object recognition," 2011.

[15] Biswas Soma, Bowyer Kevin W., and Flynn Patrick J, "Multidimensional scaling for matching low-resolution face images," 2012, vol. 34, pp. 2019–2030.

[16] R. Gopalan, S. Taheri, P. K. Turaga, and R. Chellappa, "A blur-robust descriptor with applications to face recognition," 2012, vol. 34, pp. 1220–1226.

[17] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cats visual cortex," 1962, vol. 160, p. 106.

[18] K.C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," 2005, vol. 27, pp. 684–698.

[19] R. Gross, I. Matthews, J. Cohn, T. Kanade, , and S. Baker, "Guide to the cmu multi-pie database," 2007.

[20] Ojala T, Pietikinen M, and Menp T, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," 2002, vol. 24, pp. 971–987.