2014

Affine-invariant scene categorization

Xue Wei
University of Wollongong, xw158@uowmail.edu.au

Son Lam Phung
University of Wollongong, phung@uow.edu.au

Abdesselam Bouzerdoum
University of Wollongong, bouzer@uow.edu.au

Publication Details

X. Wei, S. Lam. Phung & A. Bouzerdoum, "Affine-invariant scene categorization," in IEEE International Conference on Image Processing, 2014, pp. 1031-1035.

Affine-invariant scene categorization

Abstract
This paper presents a scene categorization method that is invariant to affine transformations. We propose a new moment-based normalization algorithm to generate an output image that is independent of the position, rotation, shear, and scale of the input image. In the proposed approach, an affine transform matrix is determined subject to the normalized image satisfying a set of moment constraints. After image normalization, a dense set of local features is extracted using scattering transform, and the global features are then formed via a sparse coding method. We evaluate the proposed method and other state-of-the-art algorithms on a benchmark dataset. The experimental results show that for images distorted with affine transformations, the proposed normalization increases the classification rate by about 28%, compared with the scene categorization approach that uses no normalization.

Keywords
categorization, scene, invariant, affine

Disciplines
Engineering | Science and Technology Studies

Publication Details
X. Wei, S. Lam. Phung & A. Bouzerdoum, "Affine-invariant scene categorization," in IEEE International Conference on Image Processing, 2014, pp. 1031-1035.

AFFINE-INVARIANT SCENE CATEGORIZATION

Xue Wei, Son Lam Phung, and Abdesselam Bouzerdoum

School of Electrical, Computer and Telecommunication Engineering
University of Wollongong, NSW, Australia, 2522
Email: xw158@uowmail.edu.au, phung@uow.edu.au, and bouzer@uow.edu.au

ABSTRACT

This paper presents a scene categorization method that is invariant to affine transformations. We propose a new moment-based normalization algorithm to generate an output image that is independent of the position, rotation, shear, and scale of the input image. In the proposed approach, an affine transform matrix is determined subject to the normalized image satisfying a set of moment constraints. After image normalization, a dense set of local features is extracted using scattering transform, and the global features are then formed via a sparse coding method. We evaluate the proposed method and other state-of-the-art algorithms on a benchmark dataset. The experimental results show that for images distorted with affine transformations, the proposed normalization increases the classification rate by about 28%, compared with the scene categorization approach that uses no normalization.

Index Terms— scene categorization, affine normalization, image moments, scattering transform.

1. INTRODUCTION

Humans can instantly grasp the conceptual meaning of an image. With a single glance, we can determine whether the scene is a beach or a forest. Gist recognition or scene categorization enables us to focus only on essential scene elements and make timely decisions that are vital to our survival. Similarly, computational methods for scene categorization can be used to rapidly provide cues about the presence of objects, or search for images and videos, and as such they are highly useful for applications in surveillance [1], navigation [2], and multimedia [3].

In recent years, researchers have studied scene categorization using several benchmark datasets, such as the 15-scene dataset [4] and the SUN dataset [5]. However, each scene in these datasets is usually captured from only a few viewing angles. In the real world, the scene could be imaged from different views, causing image variations that must be addressed by a scene categorization system. In this paper, we aim to improve the robustness of scene categorization by removing the distortions caused by arbitrary affine transformations. Because the (general) projective transformation can be locally approximated by affine transformations [6], the approach presented here constitutes a step towards developing a view-invariant scene categorization system.

The existing approaches to affine invariance can be divided into three categories: invariance by training, invariance by image normalization, and invariance by feature extraction. In the *invariance by training*, images used for training contain not only the original images but also their rotated, scaled, blurred, and deformed versions. In [7], invariant support vector machines were trained by transforming training samples with different scales, rotations, and line thicknesses. In [8], a rotation-invariant face/non-face classifier was developed by training on a large number of rotated face patterns. These techniques could easily be applied to scene categorization. However, brute force training can be time consuming, and if the training set is not carefully designed, the classifier may not learn the desired invariance.

In the *invariance by image normalization*, an input image is normalized before it is classified. Rothe *et al.* proposed the moment-based normalization that decomposes the unknown affine transformation into skew, nonuniform scaling, and rotation [9]. Zhang *et al.* studied the ambiguities of the moment-based affine normalization, and proposed a method to form a consistent normalized image [10]. Suk and Flusser decomposed the affine transformations and normalized images using low-order moments [11]. Affine normalization has been used for other applications, such as image watermarking [12] and handwritten character recognition [13].

In the *invariance by feature extraction*, the objects are described by features that are insensitive to a particular deformation. Lowe proposed the SIFT descriptor to extract scale and rotation invariant features [14]. The SIFT feature has been used for scene categorization in [15, 16]. Global feature formation methods, such as PCA [17], sparse coding [16], histogram processing [18, 19], and low-rank coding [20] reduce the sensitivity of features to geometric transformations.

In this paper, we propose a scene categorization method that is invariant to affine transformations. An input image is first transformed to a standardized image. The local features are then extracted using the scattering transform [21], and the global features are formed by sparse coding (ScSPM) [16]. The class of the image is finally determined using an SVM

classifier.

The rest of the paper is structured as follows. Section 2 presents the moment-based normalization and the sparse scattering features. Section 3 analyzes the results of experimental evaluations on a benchmark dataset, and Section 4 gives the concluding remarks.

2. PROPOSED METHOD

The proposed scene categorization approach has two major components: the image normalization and the feature extraction. Accordingly, Section 2.1 describes the moment-based normalization algorithm, whereas Section 2.2 presents the features used for scene categorization.

2.1. Moment-based affine normalization

Consider an affine transformation that is described by a transform matrix with six free parameters:

$$A = \begin{pmatrix} t_1 & t_2 & 0 \\ t_3 & t_4 & 0 \\ t_5 & t_6 & 1 \end{pmatrix}. \tag{1}$$

A pixel position in an input image $I(x, y)$ is mapped to a pixel position in the output image $I'(x', y')$ as

$$[x'\ y'\ 1] = [x\ y\ 1]\ A. \tag{2}$$

Each image can be represented by a set of moments. Given an image $I(x, y)$, its geometric moment $m_{p,q}$ of order (p, q) is defined as

$$m_{p,q} = \iint_{\Gamma} x^p y^q I(x, y) dx dy, \tag{3}$$

where Γ denotes the support of the image. The normalized geometric moment is defined as

$$\nu_{p,q} = \frac{m_{p,q}}{m_{0,0}}. \tag{4}$$

The geometric moments are not invariant to translation. To achieve translation-invariance, the central moment $\mu_{p,q}$ is defined as

$$\mu_{p,q} = \iint_{\Gamma} (x - \overline{x})^p (y - \overline{y})^q I(x, y) dx dy, \tag{5}$$

where $\overline{x} = \nu_{1,0}$ and $\overline{y} = \nu_{0,1}$. The normalized central moment is defined as

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}}. \tag{6}$$

We can prove the following result. Under the affine transformation described by matrix A, the moments $\nu'_{p,q}$ and $\eta'_{p,q}$ of the output image $I'(x', y')$ are related to the moments of the input image $I(x, y)$ as

$$\nu'_{p,q} = p!q! \sum_{(i,j)\in S_p} \sum_{(k,l)\in S_q} \frac{t_1^i\ t_2^k\ t_3^j\ t_4^l\ t_5^{p-i-j}\ t_6^{q-k-l}}{i!\ j!\ k!\ l!\ (p-i-j)!\ (q-k-l)!}\ \nu_{i+k,j+l}, \tag{7}$$

$$\eta'_{p,q} = p!q! \sum_{i=0}^{p} \sum_{j=0}^{q} \frac{t_1^i\ t_2^j\ t_3^{p-i}\ t_4^{q-j}}{i!j!(p-i)!(q-j)!}\ \eta_{i+j,p+q-i-j}, \tag{8}$$

where $S_k = \{(u, v) \in \mathbf{N}^2 \mid u + v \le k\}$.

To normalize an input image $I(x, y)$, we determine the transform matrix A so that the output image $I'(x', y')$ satisfies a set of moment constraints. In this paper, the eight constraints on the moments of the output image are specified as

$$\begin{cases} \eta'_{3,0} + \eta'_{1,2} = 0, \eta'_{1,1} = 0, & \eta'_{2,0} = c, \\ \eta'_{0,2} = c, & \nu'_{1,0} = 0, \ \nu'_{0,1} = 0, \\ \eta'_{2,1} \ge 0, & \eta'_{1,2} \ge 0. \end{cases} \tag{9}$$

Here, c is a constant used to control the size of the normalized image. A larger c produces a larger normalized image. Using Eq. (7) and (8), we can expand the eight constraints as follows:

$$\begin{cases} (t_1^3 + t_1 t_2^2)\eta_{3,0} + (t_3^3 + t_3 t_4^2)\eta_{0,3} \\ \quad + (t_1 t_4^2 + 2t_2 t_3 t_4 + 3t_1 t_3^2)\eta_{1,2} \\ \quad + (t_2^2 t_3 + 2t_1 t_2 t_4 + 3t_1^2 t_3)\eta_{2,1} = 0, \\ (t_3 t_4 \eta_{0,2} + (t_2 t_3 + t_1 t_4)\eta_{1,1} + t_1 t_2 \eta_{2,0} = 0, \\ t_3^2 \eta_{0,2} + 2t_1 t_3 \eta_{1,1} + t_1^2 \eta_{2,0} = c, \\ t_4^2 \eta_{0,2} + 2t_2 t_4 \eta_{1,1} + t_2^2 \eta_{2,0} = c, \\ t_5 + t_1 \nu_{1,0} + t_3 \nu_{0,1} = 0, \\ t_6 + t_2 \nu_{1,0} + t_4 \nu_{0,1} = 0, \\ t_1 t_2^2 \eta_{3,0} + t_3 t_4^2 \eta_{0,3} + (t_2^2 t_3 + 2t_1 t_2 t_4)\eta_{2,1} \\ \quad + (2t_2 t_3 t_4 + t_1 t_4^2)\eta_{1,2} \ge 0, \\ t_1^2 t_2 \eta_{3,0} + t_3^2 t_4 \eta_{0,3} + (t_1^2 t_4 + 2t_1 t_2 t_3)\eta_{2,1} \\ \quad + (2t_1 t_3 t_4 + t_2 t_3^2)\eta_{1,2} \ge 0. \end{cases} \tag{10}$$

The real-valued solutions of Eq. (10) form the transform matrix A. Figure 1 shows an example of the moment-based image normalization. We can see that the original image (Fig. 1(a)) and its distorted version (Fig. 1(b)) lead to the same normalized image (Fig. 1(c)).

However, for scene categorization, two different input images from the same category can produce two output images that have different orientations. Therefore, we propose an alignment step based on image moments. The normalized image is rotated through an angle $\widehat{\theta}$ which maximizes the normalized central moment $\eta'_{p,q}$:

$$\widehat{\theta} = \arg\max_{\theta} \eta'_{p,q}(\theta). \tag{11}$$

In our approach, the normalized central moment $\eta'_{2,2}$ is used. Figure 1(d) shows the moment $\eta'_{2,2}$ when the angle θ varies from 0° to 90°. The moment $\eta'_{2,2}$ is maximized at $\widehat{\theta} = 35^\circ$. Figure 1(e) shows the aligned image. Collectively, a normalized image is obtained from the input image through the transform matrix $A^* = AA_r$, where

$$A_r = \begin{pmatrix} \cos\widehat{\theta} & \sin\widehat{\theta} & 0 \\ -\sin\widehat{\theta} & \cos\widehat{\theta} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (12)$$

Finally, the normalized image is rotated by 90°, 180°, and 270° to form four output images. Examples of the four output images are shown in Fig. 2.

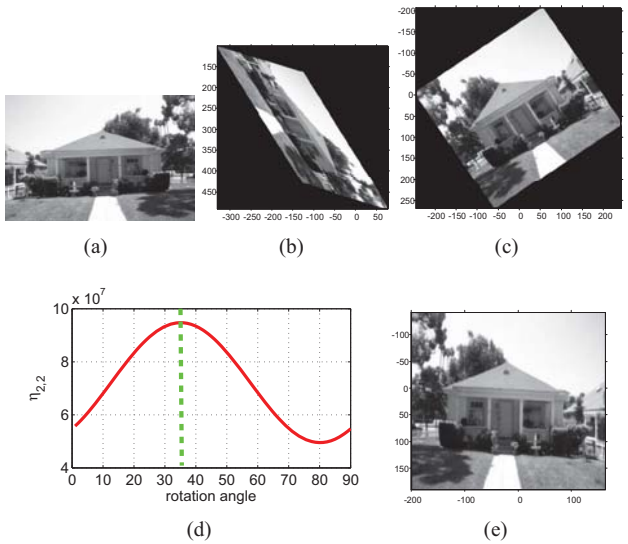

Fig. 1. An example of the proposed image normalization algorithm: (a) original input image, (b) affine-distorted image, (c) normalized image of (a) and (b), (d) normalized central moment $\eta'_{2,2}$ with θ changing from 0° to 90°, (e) aligned image where $\eta'_{2,2}$ is maximized.

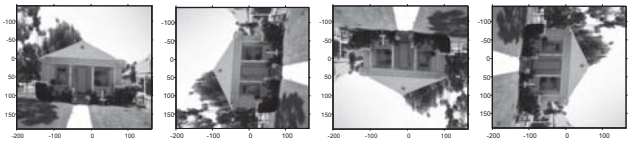

Fig. 2. The normalized images rotated through angles 0°, 90°, 180°, and 270°.

We highlight two improvements of the proposed normalization algorithm compared to the existing methods. First, our algorithm estimates the affine transformation parameters from input image moments, without decomposing the transform matrix. This strategy improves the efficiency of the normalization and avoids re-sampling errors. Second, an additional alignment step is proposed to reduce orientation uncertainty for scene categorization.

2.2. Invariant feature extraction using scattering-ScSPM

Features are extracted from the normalized images using the bag-of-word model with three stages: local feature extraction, dictionary training, and global feature coding. Several local descriptors can be applied, such as SIFT [14], HOG [22], and GIST [23]. In this paper, we adopt the scattering transform [21] to find local features.

The scattering features are stable to deformations and preserve high-frequency information that is vital for classification [21]. Given an image patch **P**, the scattering transform operators W_m are first computed based on the size of the image patch and parameters of wavelet transforms. In each layer, the operator W_m transforms the input into two new layers: an invariant layer S_{m-1} and a covariant layer U_m. Layer S_{m-1} created with low-pass filter ϕ corresponds to the energy averaging at the largest scale. Layer U_m created with band-pass filters ψ corresponds to the energy scattering along all scales and rotations. Figure 3 illustrates a second-order scattering transform. The scattering features are accumulated as

$$S = (S_0, S_1, S_2, ..., S_{m-1}). \quad (13)$$

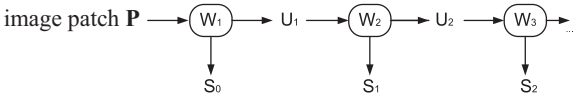

Fig. 3. A scattering representation is calculated using a cascade of wavelet-modulus operators W_m.

After extracting the local features, we build a feature dictionary based on k-means clustering. The sparse coding based spatial pyramid matching (ScSPM) [16] is then applied to convert the scattering features to global features. After feature extraction, we apply the *one-vs-all* SVM with the RBF kernel to classify the scene images. Each test image generates four normalized outputs at angles of 0°, 90°, 180°, and 270°; this strategy improves the robustness of scene categorization to various rotations. The final classification result is calculated from the mean classification score of the four normalized images.

3. EXPERIMENTAL EVALUATION AND RESULTS

In this section, we evaluate the performance of the proposed scene categorization algorithm on a public benchmark dataset.

3.1. Experimental methods

The 15-scene dataset was used to evaluate the scene categorization performance [4, 15]. It contains 10 categories of outdoor scenes and 5 categories of indoor scenes. The distorted

15-scene dataset was formed by applying random affine transformations on the original images. After applying the proposed normalization on the original or distorted images, we obtained the normalized 15-scene images.

The accuracy of normalization was evaluated by comparing the normalized form of the original image and the normalized form of the distorted image. The normalization accuracy (*NA*) is defined as the percentage of zero pixels in the difference-image. A larger *NA* value implies a more precise normalization.

In our experiment, the scattering features are calculated from overlapped patches on a dense grid. The size of the patch is 16×16 pixels. The distance between adjacent patch centers is 8 pixels. To build the dictionary, we randomly select 50 scattering features from each image as training samples. The length of the dictionary is 1024. The global features are formed by the max pooling at three spatial levels.

Two state-of-the-art feature extraction methods, GIST [23] and SIFT-ScSPM [16], were compared with the proposed method. These two descriptors were implemented in the same framework as the proposed method. The 5-fold cross validation technique was applied to measure the classification rate (*CR*). The training parameters of the RBF-SVM were determined dynamically using a separate validation set in each fold. The average value of *CR* was calculated over the five folds. The standard deviation of *CR* over the five folds was used to measure the variation in the classification performance.

3.2. Analysis of scene categorization

We first compared the proposed normalization with the decomposed normalization method in [12]. The existing method decomposes the normalization into three affine transformation matrices. As shown in Table 1, the proposed normalization is twice faster than the decomposed normalization. Moreover, the processing-time variation for the decomposed normalization is about 5 times higher than the proposed method. These results mean the proposed normalization is more efficient and stable for different affine distortions. The *NA* of the proposed method (79.4%) is also higher than that of the decomposed normalization (48.2%). The re-sampling errors caused by multiple transformations in the decomposed normalization decrease its *NA*.

Table 1. Performance of moment-based normalization.

Normalization method	*Processing time* (s)	*Normalization accuracy* (%)
Decomposed normalization [12]	1.3 ± 1.0	48.2 ± 2.8
Proposed normalization	0.6 ± 0.2	79.4 ± 2.9

Next, we evaluated the scene categorization algorithms on the original 15-scene dataset. Results in Table 2 indicate that the proposed normalization improves the classification rates of the three descriptors slightly. The adopted feature extraction method (scattering-ScSPM) outperforms the other two feature extraction method (SIFT-ScSPM and GIST).

We also evaluated the scene categorization algorithms on the distorted 15-scene dataset; the results are shown in Table 3. Several observations can be made. First, without image normalization (Table 3, Column 2), the average *CR* of the three algorithms (GIST, SIFT-ScSPM, and Scattering-ScSPM) on the distorted images is 26.6% lower than the average *CR* on the non-distorted images (Table 2, Column 2). This means affine distortions have severe effect on the existing scene categorization algorithms.

Second, on the distorted dataset, image normalization leads to higher classification rates than without image normalization; this applies to all three algorithms (GIST, SIFT-ScSPM, and Scattering-SCSPM). For example, the GIST method has a *CR* of 47.8% without image normalization, and a *CR* of 73.7% with image normalization. The SIFT-ScSPM algorithm has a *CR* of 55.3% without image normalization, and a *CR* of 84.4% with image normalization. With the proposed normalization, the average *CR* in Table 3, Column 3 was improved by 28.1%, compared with the average *CR* in Table 3, Column 2.

Third, with the proposed normalization, there are only small differences between the classification rates on the original dataset (Table 2, Column 3) and the distorted dataset (Table 3, Column 3). It shows that the proposed image normalization makes scene categorization more robust to affine distortions. Finally, the scattering-ScSPM achieves the highest classification rate 84.8% among the three descriptors.

Table 2. Scene categorization results on original images.

Feature descriptor	*Without image normalization* (%)	*With proposed normalization* (%)
GIST	73.2 ± 1.5	76.3 ± 5.5
SIFT-ScSPM	83.8 ± 1.0	84.1 ± 1.2
Scattering-ScSPM	84.0 ± 1.3	84.9 ± 1.8

Table 3. Scene categorization results on distorted images.

Feature descriptor	*Without image normalization* (%)	*With proposed normalization* (%)
GIST	47.8 ± 2.0	73.7 ± 5.3
SIFT-ScSPM	55.3 ± 1.8	84.4 ± 3.3
Scattering-ScSPM	55.5 ± 2.0	84.8 ± 1.8

4. CONCLUSION

In this paper, a new approach for scene categorization that is invariant to affine transformations is presented. An image normalization approach was proposed based on image moment constraints. Our experiments show that the proposed normalization makes scene categorization more robust to affine distortions. We believe that the proposed normalization can improve the performance not only for scene categorization, but also for other applications, like face recognition and image watermarking.

5. REFERENCES

[1] Z. Zhang, M. Li, K. Huang, and T. Tan, "Robust automated ground plane rectification based on moving vehicles for traffic scene surveillance," in *IEEE International Conference on Image Processing*, 2008, pp. 1364–1367.

[2] C. K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation and localization using gist and saliency," in *IEEE-RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4147–4154.

[3] E. Maggio and A. Cavallaro, "Learning scene context for multiple object tracking," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1873–1884, 2009.

[4] S. Lazebnik, F. Li, and A. Oliva, "Fifteen scene categories," 2006, $http://www-cvr.ai.uiuc.edu/ponce\_grp/data$.

[5] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[6] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1233–1240.

[7] D. Decoste and B. Schlkopf, "Training invariant support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 161–190, 2002.

[8] F. H. C. Tivive and A. Bouzerdoum, "A hierarchical learning network for face detection with in-plane rotation," *Neurocomputing*, vol. 71, pp. 3253–3263, 2008.

[9] I. Rothe, H. Susse, and K. Voss, "The method of normalization to determine invariants," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 366–376, 1996.

[10] Y. Zhang, C. Wen, Y. Zhang, and Y. C. Soh, "On the choice of consistent canonical form during moment normalization," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3205–3215, 2003.

[11] T. Suk and J. Flusser, "Affine normalization of symmetric objects," in *Advanced Concepts for Intelligent Vision Systems*, vol. 3708 of *Lecture Notes in Computer Science*, pp. 100–107. Springer, 2005.

[12] P. Dong, J. G. Brankov, N. P. Galatsanos, Y. Yang, and F. Davoine, "Digital watermarking robust to geometric distortions," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2140–2150, 2005.

[13] T. Wakahara and K. Odaka, "Adaptive normalization of handwritten characters using global/local affine transformation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1332–1341, 1998.

[14] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169–2178.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[17] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 506–513.

[18] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[19] W. Shao, G. Naghdy, and S. L. Phung, "Automatic image annotation for semantic image retrieval," in *Lecture Notes in Computer Science*, vol. 4781, pp. 369–378. Springer, 2007.

[20] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *IEEE International Conference on Computer Vision*, 2013, pp. 281–288.

[21] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[23] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.