Presented at: ICASSP-05, Philadelphia, PA, USA, March 19-23, 2005
Published in: Proc. IEEE Inter. Conf. Acoustic, Speech & Signal Processing
Volume II, Pages 573-576, March 2005

TRANSLATION INSENSITIVE IMAGE SIMILARITY IN COMPLEX WAVELET DOMAIN

Zhou Wang and Eero P. Simoncelli

Lab for Computational Vision, New York University, New York, NY 10003
Email: zhouwang@ieee.org, eero.simoncelli@nyu.edu

ABSTRACT

We propose a complex wavelet domain image similarity measure, which is simultaneously insensitive to luminance change, contrast change and spatial translation. The key idea is to make use of the fact that these image distortions lead to consistent magnitude and/or phase changes of local wavelet coefficients. Since small scaling and rotation of images can be locally approximated by translation, the proposed measure also shows robustness to spatial scaling and rotation when these geometric distortions are small relative to the size of the wavelet filters. Compared with previous methods, the proposed measure is computationally efficient, and can evaluate the similarity of two images without a precise registration process at the front end.

1. INTRODUCTION

Image similarity measurement is a fundamental issue in many real-world applications. For example, a *perceputal* image similarity measure can be used to estimate perceived image quality, by measuring the similarity between a distorted image and a reference image that is assumed to have perfect quality. Perhaps the simplest way to quantify the similarity between two images is the mean squared error (MSE), which is appealing because it is easy to compute and is mathematically convenient in the context of optimization. However, it has been shown that they perform poorly in image quality assessment and pattern recognition tasks (e.g., [1, 2]). We demonstrate this in Fig. 1, in which Images (b)-(g) have almost the same MSE values with respect to the reference image (a), but exhibit dramatically different perceptual quality as well as recognizability of detailed image structures. In addition, small geometrical distortions (Images (h)-(l)) can easily create much higher MSE, but do not appear to have severe degradation of image quality.

Note that all the high quality images in Fig. 1 (Images (b), (c), (h)-(l)) are associated with certain simple parametric distortions. A successful image similarity measure must be insensitive to these specific distortions. There are generally two approaches to accomplish this. The first approach, which we refer to as the "registration approach", attempts to eliminate simple parametric distortions by estimating their parameters and applying an appropriate inverse transformation to the distorted image. The second approach, which we refer to as the "invariance approach", attempts to discount specific distortions by comparing the responses of a set of measurements that are invariant to those distortions.

In this paper, we propose a new image similarity measure that does not require a precise registration process in the front, and naturally combines a number of invariants into one simple measurement. This work is inspired by the success of the spatial domain structural similarity (SSIM) index algorithm [1]. The fundamental principle of the structural approach is that the human visual system is highly adapted to extract structural information (the structures of the objects) from the visual scene, and therefore a measurement of structural similarity (or distortion) should provide a good approximation of perceptual image quality. It has been shown that a very simple SSIM algorithm provides surprisingly good image quality prediction performance for a wide variety of image distortions [1].

A major drawback of the spatial domain SSIM algorithm is that it is highly sensitive to translation, scaling and rotation of images, as demonstrated in Images (h)-(l) of Fig. 1. In this paper, we attempt to extend the current SSIM method to the complex wavelet transform domain and make it insensitive to these "non-structured" image distortions that are typically caused by the movement of the image acquisition devices, rather than the changes of the structures of the objects in the visual scene. In addition, the proposed measure shows some interesting connections with some recent computational models of biological vision (see Discussion section).

2. IMAGE SIMILARITY MEASURE

Here we consider symmetric complex wavelets whose "mother wavelets" can be written as a modulation of a low-pass filter $w(u)$= $g(u)\, e^{j\omega_c u}$, where ω_c is the center frequency of the modulated band-pass filter, and $g(u)$ is a slowly varying and symmetric function. The family of wavelets are dilated/contracted and translated versions of the mother wavelet:

$$w_{s,p}(u) = \frac{1}{\sqrt{s}}\, w\left(\frac{u-p}{s}\right) = \frac{1}{\sqrt{s}}\, g\left(\frac{u-p}{s}\right)\, e^{j\omega_c(u-p)/s}\,, \tag{1}$$

where $s \in R^+$ is the scale factor, and $p \in R$ is the translation factor. It can be shown that the continuous wavelet transform of a given real signal $x(u)$ can be written as [3]:

$$X(s,p) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega p}\, d\omega\,, \tag{2}$$

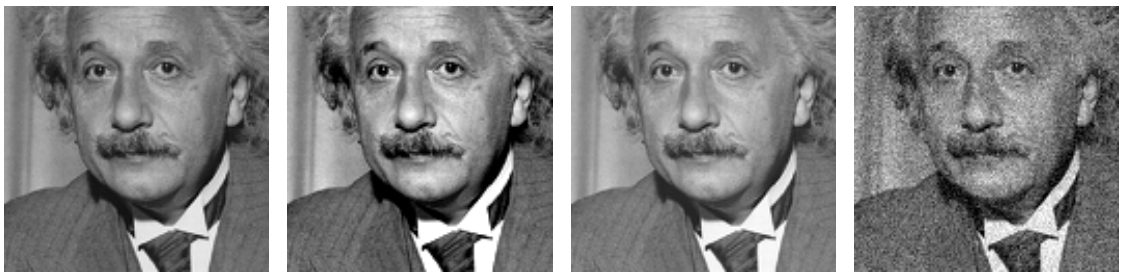
where $X(\omega)$ and $G(\omega)$ are the Fourier transforms of $x(u)$ and $g(u)$, respectively. The discrete wavelet coefficients are sampled versions of the continuous wavelet transform.

2.1. The SSIM Index

In spatial domain, the SSIM index between two image patches $\mathbf{x} = \{x_i | i = 1, ...M\}$ and $\mathbf{y} = \{y_i | i = 1, ..., M\}$ is defined as [1]

$$S(\mathbf{x}, \mathbf{y}) = \frac{(2\,\mu_x\,\mu_y + C_1)\,(2\,\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\,(\sigma_x^2 + \sigma_y^2 + C_2)}\,, \tag{3}$$

where C_1 and C_2 are two small positive constants (see [1] for details), and $\mu_x = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\sigma_x^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \mu_x)^2$ and

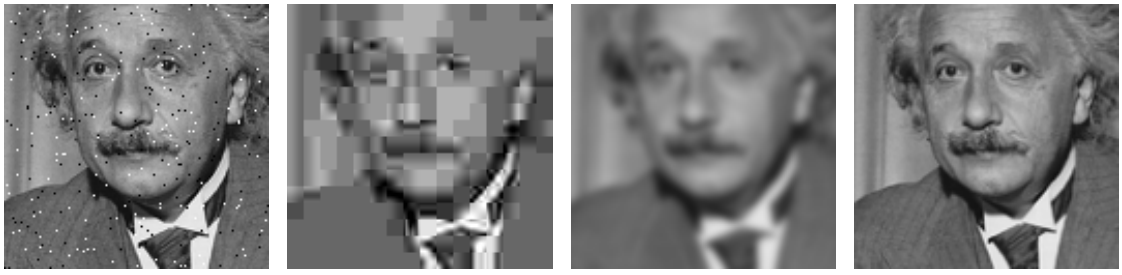


(a) MSE=0, SSIM=1
CW-SSIM=1

(b) MSE=306, SSIM=0.928
CW-SSIM=0.938

(c) MSE=309, SSIM=0.987
CW-SSIM=1.000

(d) MSE=309, SSIM=0.576
CW-SSIM=0.814

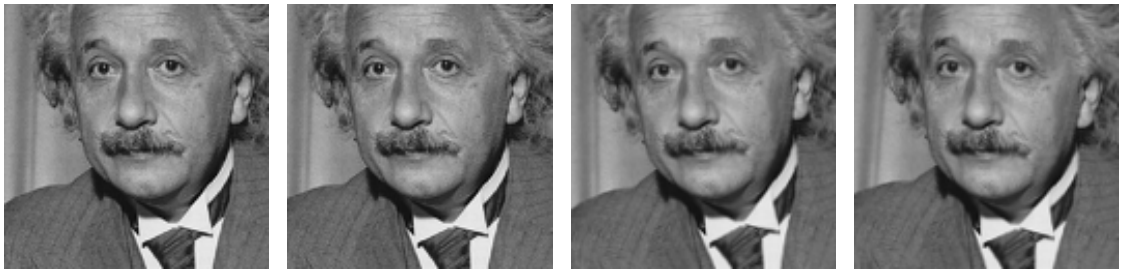


(e) MSE=313, SSIM=0.730
CW-SSIM=0.811

(f) MSE=309, SSIM=0.580
CW-SSIM=0.633

(g) MSE=308, SSIM=0.641
CW-SSIM=0.603

(h) MSE=694, SSIM=0.505
CW-SSIM=0.925



(i) MSE=871, SSIM=0.404
CW-SSIM=0.933

(j) MSE=873, SSIM=0.399
CW-SSIM=0.933

(k) MSE=590, SSIM=0.549
CW-SSIM=0.917

(l) MSE=577, SSIM=0.551
CW-SSIM=0.916

Fig. 1. Comparison of image similarity measures for images with different types of distortions. (a) reference image (8bits/pixel, assumed to have perfect quality); (b) contrast stretch; (c) mean luminance shift; (d) Gaussian noise contamination; (e) impulsive noise contamination; (f) JPEG compression; (g) blurring; (h) spatial scaling (zooming out); (i) spatial translation (to the right); (j) spatial translation (to the left); (k) rotation (counterclockwise); (l) roation (clockwise). Images are cropped from 256×256 to 128×128 for visibility. Note that Images (b)-(g) have almost the same MSE values but drastically different visual quality, which is better predicted by SSIM and CW-SSIM. Also note that MSE and SSIM are both sensitive to translation, scaling and rotation (Images (h)-(l)), and CW-SSIM is robust to these distortions.

$\sigma_{xy} = \frac{1}{M}\sum_{i=1}^{M}(x_i-\mu_x)(y_i-\mu_y)$, respectively. It can be shown that the maximum SSIM index value 1 is achieved if and only if $\mathbf{x}$ and $\mathbf{y}$ are identical.

In the complex wavelet transform domain, suppose $\mathbf{c}_x = \{c_{x,i} | i = 1, ..., N\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, ..., N\}$ are two sets of coefficients extracted at the same spatial location in the same wavelet subbands of the two images being compared, respectively. We extend the spatial domain SSIM algorithm into a complex wavelet SSIM (CW-SSIM) index (note that the coefficients are zero mean, due to the bandpass nature of the wavelet filters):

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2\left|\sum_{i=1}^{N} c_{x,i}\, c_{y,i}^*\right| + K}{\sum_{i=1}^{N}|c_{x,i}|^2 + \sum_{i=1}^{N}|c_{y,i}|^2 + K}. \qquad (4)$$

Here c^* denotes the complex conjugate of c and K is a small positive constant.

To better understand the CW-SSIM index, we rewrite it as a product of two components:

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2\sum_{i=1}^{N}|c_{x,i}||c_{y,i}| + K}{\sum_{i=1}^{N}|c_{x,i}|^2 + \sum_{i=1}^{N}|c_{y,i}|^2 + K} \cdot \frac{2\left|\sum_{i=1}^{N} c_{x,i}\, c_{y,i}^*\right| + K}{2\sum_{i=1}^{N}|c_{x,i}\, c_{y,i}^*| + K}. \qquad (5)$$

The first component is completely determined by the magnitudes of the coefficients and the maximum value 1 is achieved if and only $|c_{x,i}| = |c_{y,i}|$ for all i's. The second component, on the other hand, is fully determined by the consistency of phase changes between $\mathbf{c}_x$ and $\mathbf{c}_y$. It achieves the maximum value 1 when the phase difference between $c_{x,i}$ and $c_{y,i}$ is a constant for all i's. We consider this component as a useful measure of image structural similarity based on the believes that

- The structural information of local image features is mainly contained in the relative phase patterns of the wavelet coefficients.
- Consistent phase shift of all coefficients does not change the structure of the local image feature.

In previous work, similar phase correlation idea had been employed for image alignment [4], feature localization [5], texture description [6] and blur detection [3], but has not been used for image similarity measurement.

2.2. Sensitivity Analysis

In all the analysis below, we assume that $\mathbf{x}$ corresponds to a reference image and $\mathbf{y}$ is an altered version of the image whose similarity to the reference image is being evaluated.

Luminance and contrast changes can be roughly described as a point-wise linear transform of local pixel intensities: $y_i = a\,x_i + b$ for all i's. Due to the linear and bandpass nature of the wavelet transform, the effect in the wavelet domain is a constant scaling of all the coefficients, i.e., $c_{y,i} = a\,c_{x,i}$ for all i's. Substitute this into Eq. (5), we can see that a perfect value 1 is obtained for the second component and the first component gives

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2a + K/\sum_{i=1}^{N}|c_{x,i}|^2}{1 + a^2 + K/\sum_{i=1}^{N}|c_{x,i}|^2}\,. \tag{6}$$

At strong image features (large coefficient magnitudes), $K/\sum_{i=1}^{N}|c_{x,i}|^2$ is small and can be ignored, leading to an insensitive measure (compared with MSE) – scaling the magnitude by a factor of 10% ($a = 1.1$) only causes reduction of the SSIM value from 1 to 0.9955. The measure is even less sensitive at weaker image features (small coefficient magnitudes).

Translation, scaling and rotation in the 2-D spatial domain can be written as

$$y\begin{pmatrix}u_1\\u_2\end{pmatrix} = x\left(\begin{pmatrix}\cos\Delta\theta & -\sin\Delta\theta\\ \sin\Delta\theta & \cos\Delta\theta\end{pmatrix}\begin{pmatrix}1+\Delta s_1 & 0\\ 0 & 1+\Delta s_2\end{pmatrix}\begin{pmatrix}u_1\\u_2\end{pmatrix} + \begin{pmatrix}\Delta t_1\\ \Delta t_2\end{pmatrix}\right), \tag{7}$$

where $(1+\Delta s_1, 1+\Delta s_2)$, $\Delta\theta$, and $(\Delta t_1, \Delta t_2)$ are the scaling, rotation and translation factors, respectively. When $\Delta\theta$ is small, we have $\cos\Delta\theta \approx 1$ and $\sin\Delta\theta \approx \Delta\theta$, and therefore

$$\begin{aligned} y\begin{pmatrix}u_1\\u_2\end{pmatrix} &\approx x\begin{pmatrix}u_1 + (u_1\Delta s_1 - u_2\Delta\theta + \Delta t_1 - u_2\Delta s_2\Delta\theta)\\ u_2 + (u_2\Delta s_2 + u_1\Delta\theta + \Delta t_2 + u_1\Delta s_1\Delta\theta)\end{pmatrix}\\ &= x\begin{pmatrix}u_1 + \Delta u_1\\ u_2 + \Delta u_2\end{pmatrix}, \end{aligned} \tag{8}$$

From Eq. (8), we see that when (u_1, u_2) is not far away from the origin, small amount of translation, scaling and rotation can be locally approximated by a small translation $(\Delta u_1, \Delta u_2)$. For easy analysis, let us consider the 1-D case $y(u) = x(u + \Delta u)$. This corresponds to a linear phase shift in the Fourier domain $Y(\omega) = X(\omega)e^{j\omega\Delta u}$. Substitute this into Eq. (2), we obtain

$$\begin{aligned} Y(s,p) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\sqrt{s}\,G(s\,\omega - \omega_c)\,e^{j\omega(p+\Delta u)}\,d\omega\\ &= \frac{e^{j\omega_c\Delta u/s}}{2\pi}\int_{-\infty}^{\infty} X(\omega)\sqrt{s}\,G(s\omega - \omega_c)e^{j\omega p}e^{j(\omega-\omega_c/s)\Delta u}d\omega\\ &\approx X(s,p)\,e^{j\omega_c\Delta u/s}\,. \end{aligned} \tag{9}$$

Here the approximation is valid when Δu is small compared to the spatial extent of $g(u)$. Similar result can be obtained for the 2-D case. Consequently, the corresponding discrete wavelet coefficients $\{c_{y,i}\}$ and $\{c_{x,i}\}$ (discrete samples of $X(s,p)$ and $Y(s,p)$) are approximately phase shifted versions of each other. Therefore, based on the analysis of the CW-SSIM index in Section 2.1, we have $\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) \approx 1$, where the accuracy of the approximation depends on the magnitudes of the translation, scaling and rotation factors as well as the shape of the envelop of the wavelet filter.

3. TEST

To apply the CW-SSIM measure for comparing images, we first decompose the two given images being compared using a complex version [6] of the "steerable pyramid" transform [7] (a type of redundant wavelet transform that avoids aliasing in subbands). We then move a sliding window (of size 7×7) step by step across each wavelet subband. At each step, the CW-SSIM index is calculated within the sliding window using Eq. (4). The overall similarity of the two images is estimated using the average of the local CW-SSIM measures in all subbands (or a subset of all the subbands).

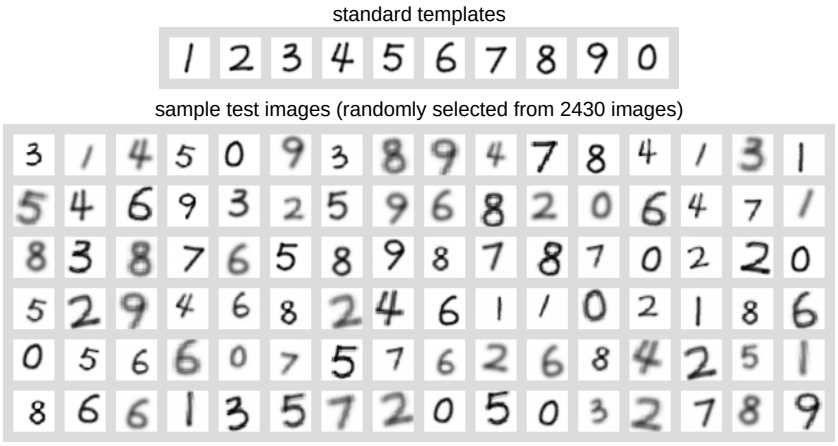
Figure 1 demonstrates the CW-SSIM measure for image quality assessment. A 2-scale, 16-orientation steerable pyramid decomposition is constructed and the 16 subbands at the second scale are used by the CW-SSIM measure. It can be seen that images with almost the same MSE values but different distortion types (Images (b)-(g)) have drastically different visual quality, which is better predicted by SSIM and CW-SSIM. However, the SSIM method fails to provide useful quality prediction when the images are slightly shifted, scaled or rotated (Images (h)-(l)). These are effectively accounted for by CW-SSIM, which gives significantly higher scores to Images (b), (c) and (h)-(l) than to Images (d)-(g).

In Fig. 2, we demonstrate the effectiveness of the CW-SSIM measure using a pattern matching test. We first manually created ten standard digit templates with a size of 32×32, as shown in Fig. 2. A total of 2430 distorted images (243 for each digit) were then generated by shifting, scaling, rotating, and blurring the standard templates (examples shown in Fig. 2). We then "recognize" each distorted image based on direct image matching with the ten standard templates, without any registration process in the front. MSE, SSIM and CW-SSIM are used as the matching standards, where CW-SSIM employs 4 subands of the second scale in a 2-scale, 4-orientation steerable pyramid transform.

The recognition performance is significantly different when different similarity measures are employed. As expected, the MSE and spatial domain SSIM measures are sensitive to translation, scaling and rotation of images, thus poor correct recognition rates were obtained. By contrast, the performance of the CW-SSIM measure is surprisingly good, achieving an overall correct recognition rate of 97.7%! It needs to be emphasized that this experiment is only a simple test of image similarity measures, but not a complete test of digit recognition systems. However, we find it impressive that this approach, which does not rely on any registration or intensity normalization preprocessing, does not include any probabilistic model for either the image patterns or the distortions, and requires no training, works as well as it does.

4. CONCLUSION AND DISCUSSION

We propose the CW-SSIM index method, an image similarity measure that is simultaneously insensitive to luminance change, con-



correct recognition rate (%)

digit	MSE	SSIM	CW-SSIM
1	84.0	76.1	100
2	65.4	45.3	98.4
3	49.4	47.7	97.1
4	63.8	41.6	100
5	47.7	18.5	96.3
6	56.4	42.0	97.9
7	68.3	60.9	94.2
8	49.8	39.1	99.6
9	59.3	51.4	100
0	51.4	46.5	93.0
all	59.6	46.9	97.7

Fig. 2. Pattern matching without registration. Each test image (from a database of 2430 images) is matched to the ten standard templates using MSE, SSIM and CW-SSIM as the similarity measures, without any normalization or registration process in the front. The test image is then "recognized" as belonging to the category that corresponds to the best similarity score. The resulting correct recognition rates show that both MSE and SSIM are sensitive to translation, scaling and rotation of images, but CW-SSIM exhibits much stronger robustness.

trast change, and small translation, scaling and rotation of images. It is computationally efficient in comparison with typical registration-based methods, which usually require a more complicated procedure to search or estimate the registration parameters.

The proposed algorithm shows some interesting connections with several computational models that have been successfully used to account for a variety of biological vision behaviors. These models include: 1) The involvement of bandpass visual channels in image pattern recognition tasks [8]; 2) Representation of phase information in primary visual cortex using quadrature pairs of localized bandpass filters [9]; 3) The computation of complex-valued product in visual cortex [10]; 4) The computation of local energy (using sums of squared responses of quadrature-pair filters) by complex cells in visual cortex [11]; and 5) Divisive normalization of filter responses (using summed energy of neighboring filter responses) in both visual and auditory neurons [12].

It is important to realize the limitations of the current algorithm. First, the CW-SSIM measure does not provide any correspondence information between the pixels of the two images being compared (a disadvantage compared to registration-based approaches). Second, the method works only when the amount of translation, scaling and rotation is small (compared to the wavelet filter size). This problem may be solved by using a multi-scale, coarse-to-fine method, with a rough registration adjustment between scales, e.g., [13]. The inclusion of invariants to additional parametric distortions may also lead to further improvement.

5. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[2] P. Y. Simard, Y. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition – tangent distance and tangent propagation," *International Journal of Imaging Systems and Technology*, vol. 11, no. 3, 2000.

[3] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in *Adv. Neural Information Processing Systems (NIPS03)*, vol. 16, (Cambridge, MA), MIT Press, May 2004.

[4] C. Kuglin and D. Hines, "The phase correlaton image alighment method," in *Proc. IEEE Int. Conf. Cybern. Soc.*, pp. 163–165, 1975.

[5] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, pp. 303–313, 1987.

[6] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l J Computer Vision*, vol. 40, pp. 49–71, 2000.

[7] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory*, vol. 38, pp. 587–607, Mar 1992.

[8] J. A. Solomon and D. G. Pelli, "The visual filter mediating letter identification," *Nature*, vol. 369, pp. 395–397, 1994.

[9] D. A. Pollen and S. F. Ronner, "Phase relationships between adjacent simple cells in the cat," *Science*, no. 212, pp. 1409–1411, 1981.

[10] I. Ohzawa, G. DeAngelis, and R. Freeman, "Stereoscopic depth descrimination in the visual cortex: Neurons ideally suited as disparity detectors," *Science*, no. 249, pp. 1037–1041, 1990.

[11] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J Optical Society of America*, vol. 2, pp. 284–299, Feb 1985.

[12] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, pp. 819–825, August 2001.

[13] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *IEEE Int. Conf. Comp. Vis. & Patt. Reco.*, pp. 310–315, June 1991.