A Mutual-Information Scale-space for Image Feature Detection and Feature-based Classification of Volumetric Brain Images

Matthew Toews and William M. Wells III
Harvard Medical School, Brigham and Women's Hospital
Boston, MA, USA. {mt,sw}@bwh.harvard.edu

Abstract

This paper proposes a novel information theoretic scale-space for salient feature detection, based on the mutual information (MI) of image measurement and location. The MI scale-space is designed to identify image regions whose measurements are maximally informative regarding spatial location. A framework for computing the MI scale-space is proposed, based on combining information theory with Gaussian scale-space theory, where uncertainty in spatial location is explicitly defined by the heat equation. Experiments investigate the use of MI features for feature-based classification of Alzheimer's subjects in volumetric magnetic resonance imagery from a public data set, where MI features result in higher classification accuracy than features selected according to the established difference-of-Gaussian (DOG) criterion [15].

1. Introduction

Medical images contain a large amount of information, and efficiently processing sizable image sets can be achieved by focusing computational resources on image features or patterns which are most salient or informative for the task at hand. A variety of mechanisms have been proposed for identifying salient image features. Biologically-motivated approaches have investigated spatial maps which encode the degree of saliency as a function of image location [12, 8]. Computer vision-based approaches have focused on detecting patterns in space and in scale which can be used to efficiently establish correspondences between different images of the same scenes or objects [10, 15, 18, 1, 17, 27]. Many approaches operate by identifying patterns which maximize an image-based saliency criterion, examples of criteria include the magnitude of Gaussian derivatives [15, 18] or image phase [1]. In the medical imaging literature, the difference-of-Gaussian (DOG) scale-space [15] has found use in a variety of tasks, including image matching [22, 2] and feature-based group analysis [26].

Information theory [23, 3] provides a principled basis for evaluating feature saliency, which can be expressed and quantified in terms of information content. Several information theory-based saliency criteria have been proposed, including the Kullback-Leibler divergence [9] and entropy [10, 24]. The drawback of most information theory-based criteria to date is that they do not effectively quantify the ability to localize features within the image, as they focus exclusively on the information content of image measurements, e.g. the entropy of image intensity distributions [10, 24]. Several authors have proposed defining saliency in terms of the mutual information between image measurements and spatial location [25, 6], however this research has not been extended to identifying natural structure at a characteristic scale for practical feature detection.

The primary contribution of this paper is a novel scale-space, based on the mutual information (MI) of image measurements and spatial location. Unlike previous information theory-based approaches based solely on the information content of image measurements [9, 24, 10], the MI scale-space explicitly quantifies the ability of image measurements to predict spatial location. High MI reflects image measurements that are highly informative regarding spatial location, and thus highly salient for tasks such as image matching or classification. The MI criterion is formulated within a Gaussian scale-space [28, 13, 14], where spatial location is defined as a Gaussian random variable. The Gaussian scale-space is generalized from scalar-valued intensities to distributions over image measurements conditioned on image location and scale. A computational framework is developed, which can be used to efficiently generate the MI scale-space for the purpose of salient feature detection.

The remainder of this paper is organized as follows. Section 2 presents previous work in scale-space theory, information theory and salient feature identification. Section 3 outlines the MI saliency criterion and framework for computing the MI criterion from a Gaussian scale-space. Section 4 details experiments comparing MI feature detection with the widely-used DOG detection criterion [15], in the context of feature-based classification of volumetric magnetic resonance (MR) images of Alzheimer's and healthy subjects from the OASIS data set [16]. Results show that MI features result in higher classification accuracy. A discussion follows in Section 5.

2. Previous Work

2.1. Scale-space Theory

Scale-space theory is concerned with describing image measurements in a manner independent of the sampling resolution. A scale-space is a function $I(x, \sigma)$ which describing image measurements I at image location x and scale σ. The Gaussian scale-space proposed by Witkin [28] is arguably the most prevalent in the computer vision literature, defined by recursive convolutions of an image with a Gaussian kernel:

$$I(x, \sigma) = \int I(X, \kappa\sigma) G(X; x, (1-\kappa)\sigma) dX, \quad (1)$$

where $G(X; x, \sigma)$ represents the Gaussian kernel of mean x and variance σ, and $0 \leq \kappa \leq 1$ is a constant that determines the multiplicative sampling rate in the scale dimension. Subsequent work by Koenderink shows that a Gaussian scale-space is consistent with modeling the image as a diffusion process governed by the heat equation [13]:

$$\partial_\sigma I(x, \sigma) = \frac{1}{2} \nabla^2 I(x, \sigma), \quad (2)$$

where $G(X; x, \sigma)$ arises as the Green's function of the partial differential equation (2). Lindeberg uses a set of axioms to show that the Gaussian kernel is the only reasonable choice in generating $I(x, \sigma)$ [14], including causality, non-creation and non-enhancement of local extrema, semigroup structure and scale-invariance. Aside from computer vision, the Gaussian scale-space is proposed as a model of image processing in biological vision systems [29].

2.2. Information Theory

Information theory is concerned with measuring the information content of random variables. Central to information theory is the notion of entropy, first introduced by Shannon [23] and since generalized by others, e.g. the Renyi entropy [3]. Consider a random variable I defined by a distribution $p(I)$ over a space of image measurements I. The information content of I is quantified by its entropy $H(I)$ defined as:

$$H(I) = \int p(I) \log \frac{1}{p(I)} dI. \quad (3)$$

The amount of information shared between two random variables I and X can be quantified by their mutual information $MI(I, X)$, defined in terms of entropies as:

$$MI(I, X) = H(I) - H(I|X), \quad (4)$$

where $H(I|X)$ is the conditional entropy of I given X. The conditional entropy quantifies the amount of information remaining regarding I when X is known, and is defined as:

$$H(I|X) = \int \int p(I, X) \log \frac{1}{p(I|X)} dI dX. \quad (5)$$

Note that the MI is upper bounded by the marginal entropy $H(I)$:

$$0 \leq MI(I, X) \leq H(I), \quad (6)$$

where $MI(I, X) = 0$ implies that I and X are statistically independent, i.e. $p(I|X) = p(I)$, and $MI(I, X) = H(I)$ implies full dependence between I and X.

2.3. Salient Feature Selection

Natural images contain a large amount of data, and efficiently performing tasks such as image matching or object recognition requires focusing computation on a reduced set of salient or informative image patterns. While early interest point detectors identified salient image points such as corners [20, 7], most image patterns have a characteristic scale which is generally unknown *a priori*, and a modern methodology is to identify salient image patterns both in scale and in space [14]. Models of biological vision propose computing saliency maps, in which the degree of saliency at each image point is computed via image filter responses computed at multiple scales [12, 8]. A body of computer vision literature focuses on detecting local image patterns in a manner invariant to geometrical deformations such as scale [15, 18] and affine [19, 17, 27] changes, where patterns correspond to extrema of a particular criterion computed from an image scale-space. A variety of criteria have been proposed, including derivatives in scale [15] and space [18], image phase [1], etc.

Information theoretic criteria provide a principled means of quantifying pattern saliency when measurements I are represented as a probability $p(I|x, \sigma)$ at a particular location x and scale σ in scale-space. Jagersandt proposes identifying patterns in scale-space where the Kullback-Leibler divergence between measurement distributions at different scales is maximized [9]:

$$D = \int p(I|x, \sigma_i) \frac{p(I|x, \sigma_i)}{p(I|x, \sigma_{i-1})} dI, \quad (7)$$

where σ_i and σ_{i-1} are adjacent scales. Sporring investigates the computation of entropy change in Gaussian scale-space [24], showing that entropy is generally an increasing function of scale which changes most rapidly at the characteristic scale of texture patterns in the image, and that derivatives with respect to scale should be normalized by parameter σ:

$$\frac{dH(I|x, \sigma)}{d\sigma} \propto \sigma(H(I|x, \sigma_i) - H(I|x, \sigma_{i-1})). \quad (8)$$

Kadir et al. propose identifying points in scale-space in which the entropy is maximal [10, 11]. They note, however, that many high-entropy features arise from unimportant textured or random image patterns, and propose weighting entropy by a heuristic factor $\int \left| \frac{d\ p(I|x,\sigma)}{d\sigma} \right| dI$ in order to better quantify the saliency of detected features.

The major drawback of most information theory-based saliency criteria is that they focus on the information content of image measurements I alone. The MI criterion presented in Section 3 addresses localizability by explicitly quantifying the information shared between image measurements and spatial location. An additional drawback is computational complexity: the time required to compute entropy-based features when $p(I|x,\sigma)$ is represented as a discrete intensity histogram [10, 11] is several orders of magnitude greater than for other detectors [21]. The primary difficulty lies in the histogram representation, which is computationally expensive for uni-dimensional greyscale measurements alone and potentially intractable for high-dimensional measurements, e.g. diffusion-weighted images, as the number of histogram bins increases exponentially with dimensionality. A mixture model-based data representation is proposed in Section 4 which casts measurements into a small number of latent classes, allowing efficient information computation in a manner independent of data dimensionality.

3. Mutual Information of Measurement and Location

The ability to localize image patterns is of key importance in a variety of image processing tasks, e.g. image matching. For this reason, we argue that salient image features should not only bear informative measurements I, but that measurements should be informative with respect to spatial location X. We thus propose a new information theory-based criterion for measuring the saliency of an image pattern at location x and scale σ, the mutual information (MI) of image measurement I and spatial location X:

$$MI(I;X|x,\sigma) = H(I|x,\sigma) - H(I|X,x,\sigma), \tag{9}$$

where high $MI(I;X|x,\sigma)$ indicates an image pattern that is both informative and localizable, and thus salient. Note that in order for MI to be high, the measurement entropy $H(I|x,\sigma)$ must be high as in entropy-based selection criteria [10], however the conditional entropy $H(I|X,x,\sigma)$ must also be low, meaning that I is informative regarding X.

The MI criterion is formulated in terms of information theory [3] in order to quantify probabilistic uncertainty in a principled manner, and in terms of scale-space theory [14] in order to identify salient regions both in space and in scale. Here, image measurements I are defined probabilistically as distributions $p(I|x,\sigma)$ conditioned on location and scale. The definition of scale-space in Equation (1) must thus be generalized from scalar-valued intensity measurements to distributions over intensities or intensity classes. We propose doing this via the following expression:

$$p(I|x,\sigma) = \int G(X;x,(1-\kappa)\sigma)p(I|X,\kappa\sigma)dX, \tag{10}$$

where the integral in Equation (10) is evaluated independently for each discrete intensity level I. This is computationally expensive for a large number of intensity levels, however in practice, I can be represented as a distribution over a small number of intensity classes as described later in Section 4.

In order to compute the MI expression in Equation (9), $H(I|x,\sigma)$ at coordinate (x,σ) can be obtained directly from scale-space $p(I|x,\sigma)$. The conditional entropy $H(I|X,x,\sigma)$ of I given random variable X is defined as:

$$H(I|X,x,\sigma) = \int\int p(I,X|x,\sigma)\log\frac{1}{p(I|X,x,\sigma)}dIdX, \tag{11}$$

and requires distributions $p(I,X|x,\sigma)$ and $p(I|X,x,\sigma)$. Suitable expressions for these distributions can be obtained as follows. The chain rule of probability results in the following expression:

$$\begin{aligned} p(I|x,\sigma) &= \int p(I,X|x,\sigma)dX, \\ &= \int p(X|x,\sigma)p(I|X,x,\sigma)dX. \end{aligned} \tag{12}$$

Assuming equivalence between the integrands in Equations (12) and (10), the following equalities are obtained:

$$p(X|x,\sigma) = G(X;x,(1-\kappa)\sigma), \tag{13}$$

$$p(I|X,x,\sigma) = p(I|X,\kappa\sigma), \tag{14}$$

$$p(I,X|x,\sigma) = G(X;x,(1-\kappa)\sigma)p(I|X,\kappa\sigma). \tag{15}$$

Note that in Equation (13), the probabilistic uncertainty in spatial location about x at scale σ is quantified by a Gaussian distribution centered on x with variance $(1-\kappa)\sigma$. Substituting Equations (15) and (14) into Equation (11), the final expression for the conditional entropy is:

$$\begin{aligned} &H(I|X,x,\sigma) \\ &= \int\int G(X;x,(1-\kappa)\sigma)p(I|X,\kappa\sigma)\log\frac{1}{p(I|X,\kappa\sigma)}dIdX, \\ &= \int G(X;x,(1-\kappa)\sigma)H(I|X,\kappa\sigma)dX. \end{aligned} \tag{16}$$

The computational framework for the MI criterion is shown in Figure 1. Note that three scale-spaces are involved, including a Gaussian scale-space, an entropy scale-space and the MI scale-space.

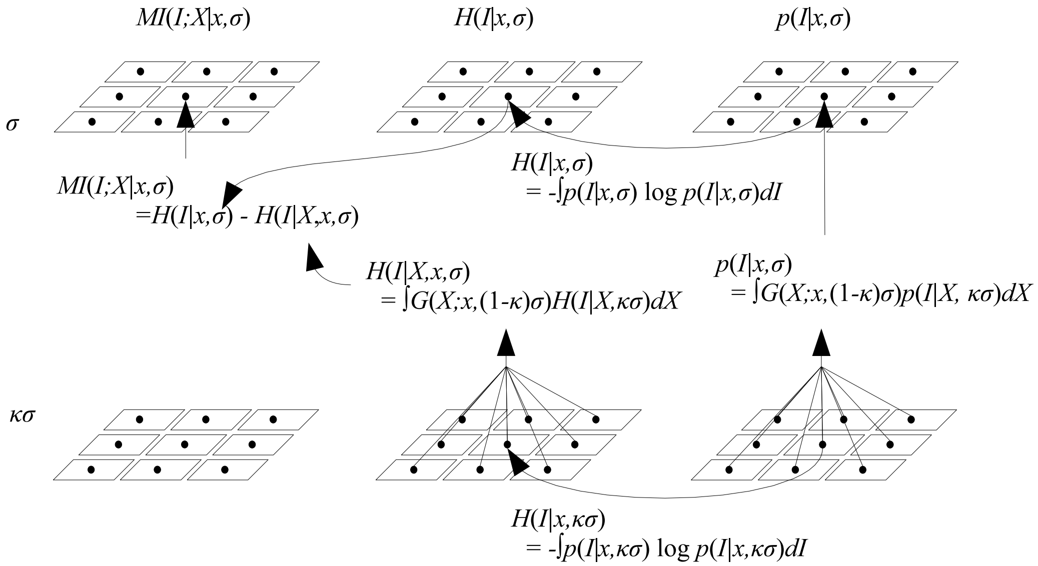

$MI(I;X|x,\sigma)$
$H(I|x,\sigma)$
$p(I|x,\sigma)$
σ
$H(I|x,\sigma)$
$= -\int p(I|x,\sigma) \log p(I|x,\sigma)dI$
$MI(I;X|x,\sigma)$
$=H(I|x,\sigma) - H(I|X,x,\sigma)$
$H(I|X,x,\sigma)$
$= \int G(X;x,(1-\kappa)\sigma)H(I|X,\kappa\sigma)dX$
$p(I|x,\sigma)$
$= \int G(X;x,(1-\kappa)\sigma)p(I|X, \kappa\sigma)dX$
$\kappa\sigma$
$H(I|x,\kappa\sigma)$
$= -\int p(I|x,\kappa\sigma) \log p(I|x,\kappa\sigma)dI$

Figure 1. The computational framework for the MI criterion. Squares represent image locations, upper and lower rows represent adjacent scales σ and $\kappa\sigma$ in a scale-space, where $\sigma \geq \kappa\sigma$. Arrows indicate the flow of data, and arrows combining multiple locations indicate integration over space X. The measurement scale-space $p(I|x,\sigma)$ (right) is a Gaussian scale-space in the style of Witkin and Koenderink. The entropy scale-space $H(I|x,\sigma)$ (center) is derived directly from $p(I|x,\sigma)$. The MI scale-space $MI(I;X|x,\sigma)$ (left) is computed from adjacent levels in the entropy scale-space.

4. Experiments

Experiments here compare the MI criterion with the DOG criterion [15] in the context of feature-based classification of Alzheimer's disease in MR images of the human brain. The OASIS dataset is used [16], consisting of T1-weighted MRIs of 100 Alzheimer's (AD) and 98 normal control (NC) subjects.

Feature Extraction: DOG and MI scale-spaces are both implemented by sampling the scale dimension 3 times per octave, in a manner similar to Lowe [15], over 4 octaves. Scale-invariant features are identified as the location and scale (x, σ) at which the particular criteria are maximal, in the case of the DOG criteria both minima and maxima are considered. Note that derivative of the MI criterion with respect to scale is normalized by multiplication by scale factor σ [24]. In the case of the MI criterion, measurements are formulated as a distribution $p(I|x, \sigma)$ at each image location and scale. $p(I|x, \sigma)$ can be represented as a multinomial distribution over a set of discrete events. Events can be defined by partitioning the measurement space into discrete histogram bins [10, 11], however this approach is computationally expensive even for unidimensional greyscale image measurements [21], and intractable for multi-dimensional measurements as the number histogram bins increases exponentially with data dimensionality. For the sake of computational efficiency, mixture modeling can be used to reduce $p(I|x, \sigma)$ to a distribution over a small set of K latent classes $C = \{C_1, \dots, C_K\}$ [4]:

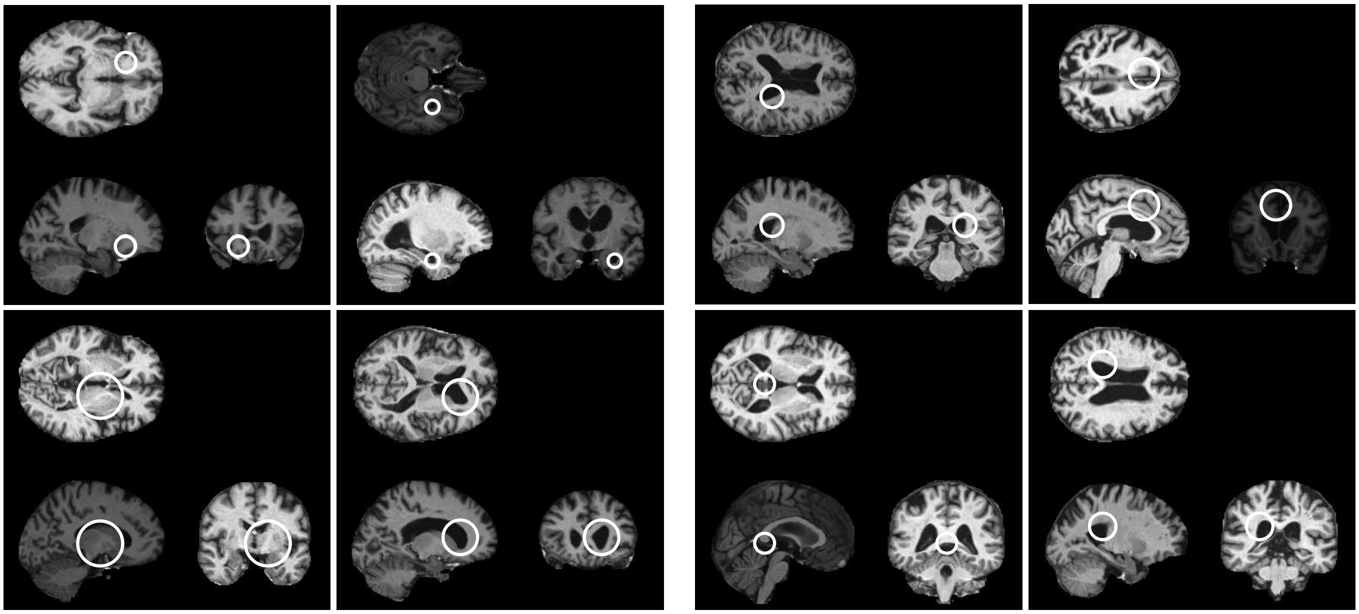
$$p(I|x, \sigma) = \sum_i^K p(I|C_i, x, \sigma)p(C_i|x, \sigma),$$

$$p(C|I, x, \sigma) \propto p(I|C, x, \sigma)p(C|x, \sigma). \tag{17}$$

The computational and space complexity of scale-space generation is thus $O(KM\ log\ M)$, where M is the number of image intensity measurements. Here a binary mixture model with $K = 2$ is used, which can be learned offline using the expectation maximization (EM) algorithm [5] and used to efficiently cast image measurements into a binomial distribution $p(C|I, x, \sigma)$. In the case of grey scale measurements, this intuitively corresponds to a distribution over 'low' and 'high' intensity classes.

Figure 2 illustrates features extracted according to the DOG and MI criteria. Qualitatively, DOG features arise from blob-like image structures, eg. ventricles or white matter, whereas MI features tend to arise from informative boundaries between brain tissues.

Classification: Feature-based classification is performed in a manner similar to the feature-based morphometry technique [26]. Extracted features are binned into a histogram defined in normalized Montreal neurological institute (MNI) brain reference space, where bins are defined geometrically in terms of location and scale. Let i represent the index of a histogram bin, and let f_i represent the binary presence/absence of a feature in bin i. Under the assumption of conditionally independent features, a Bayes



(a) DOG

(b) MI

Figure 2. Examples of features extracted according to the DOG (a) and MI (b) criteria. In each image, the location and scale of a feature extracted in a 3D MRI volume is illustrated as a white circle in coronal, sagittal and axial slices. DOG features shown in (a) arise from blob-like structures, e.g. ventricles, subcortical structures and white matter. MI features in (b) arise from patterns along the boundary between white/grey matter and cerebrospinal fluid.

ratio can be used to evaluate the posterior probability of a set of features $\{f_i\}$ extracted in a subject given either AD or NC subject groups:

$$\frac{p(AD|\{f_i\})}{p(NC|\{f_i\})} = \frac{p(AD)}{p(NC)} \prod_i \frac{p(f_i|AD)}{p(f_i|NC)}. \qquad (18)$$

In Equation (18), $\frac{p(AD)}{p(NC)}$ represents the ratio of prior probabilities of AD vs. NC subjects, and $\frac{p(f_i|AD)}{p(f_i|NC)}$ represents the likelihood ratio of AD vs. NC subjects associated with the occurrence/absence of feature f_i. Leave-one-out classification is performed on a set of N subjects, where a classifier is trained from features extracted in $N-1$ subjects by estimating likelihoods $p(f_i|AD)$ and $p(f_i|NC)$ in Equation (18), then tested by calculating the Bayes ratio for the 1 subject left out, for all each subject. The receiver operating characteristic (ROC) curves for classification are shown in Figure 3. The equal error classification rates obtained are 81.8% for DOG features and 85.7% for MI features.

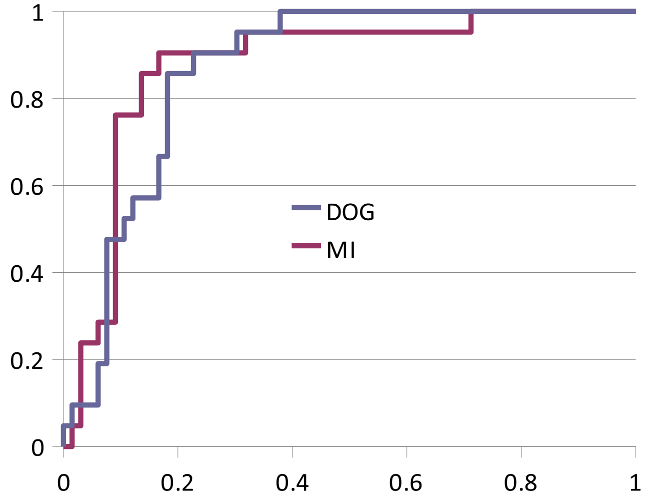


Figure 3. ROC curves comparing leave-one-out classification performance of Alzheimer's subjects, aged 60-80 years of age, where AD subjects exhibit mild dementia (CDR=1.0).

5. Discussion

This paper presents a novel scale-space for salient feature detection in volumetric medical imagery, based on the mutual information of image measurement and location, combining Gaussian scale-space and information theories. Intensity measurements are cast into a distribution over a small number of intensity classes to achieve computational tractability. Experiments compare features selected according to maximum MI are to features selected according to the popular DOG feature criterion, in volumetric MR imagery of the human brain. Qualitatively, maxima in the MI scale-space correspond to informative structure along tissue boundaries, while DOG features correspond to blob-like image structures. A quantitative comparison of MI vs.

DOG features is performed in the classification of healthy subjects vs. subjects with mild AD from the OASIS data set [16]. Using a Bayesian classifier based on conditionally independent features, MI features result in improved classification over DOG features, with equal error classification rates of 85.7% vs. 81.8%, respectively.

Various future directions exist. The MI criterion identifies regions in which image measurements are maximally informative regarding spatial location, the could prove to be effective in a feature-based registration framework. The MI scale-space is a general formulation that can be used to extract informative image features in a variety of image types, including vector-valued data such as diffusion-weighted images. While many feature selection criteria such as the DOG criterion are limited to scalar-valued images, the MI criterion could be used for feature-based analysis of vector-valued images. The computational framework for the MI scale-space could be used to implement other information theoretic saliency criteria, such as the entropy-based criterion of Kadir and Brady [10], in a manner consistent with Gaussian scale-space theory.

6. Acknowledgement

This research was supported by NIH grant P41 RR13218 and an NSERC Postdoctoral Fellowship. OASIS data involved grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

References

[1] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *CVPR*, volume 1, pages 736–743, 2003. 1, 2

[2] W. Cheung and G. Hamarneh. N-sift: N-dimensional scale invariant feature transform for matching medical images. In *ISBI*, 2007. 1

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, New York, 1991. 1, 2, 3

[4] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. An information theoretic approach for non-rigid image registration using voxel class probabilities. In *MICCAI*, pages 812–820, 2003. 4

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2nd edition, 2001. 4

[6] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. pages 1–6, 2007. 1

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988. 2

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, November 1998. 1, 2

[9] M. Jagersandt. Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach. In *ICCV*, pages 195–202, 1995. 1, 2

[10] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, 2001. 1, 3, 4, 6

[11] T. Kadir, A. Zisserman, and M. Brady. An affne invariant salient region detector. In *ECCV*, 2004. 3, 4

[12] C. Koch and S. Ullman. Shifts in selective visual attention: Toward the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985. 1, 2

[13] J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984. 1, 2

[14] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer, 1994. 1, 2, 3

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2, 4

[16] D. Marcus, T. Wang, J. Parker, J. Csernansky, J. Morris, and R. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 19:1498–1507, 2007. 2, 4, 6

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002. 1, 2

[18] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 1, 2

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005. 2

[20] H. P. Moravec. Visual mapping by a robot rover. In *Proc. of the 6th International Joint Conference on Artificial Intelligence*, pages 598–600, 1979. 2

[21] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *IJCV*, 73(3):263–284, 2007. 3, 4

[22] D. Ni, Y. Qu, X. Yang, Y. P. Chui, T.-T. Wong, S. Ho, and P. A. Heng. Volumetric ultrasound panorama based on 3D SIFT. In *MICCAI*, 2008. 1

[23] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948. 1, 2

[24] J. Sporring. The entropy of scale-space. In *ICPR*, pages 900–904, 1995. 1, 2, 4

[25] M. Toews and W. M. Wells III. Bayesian registration via local image regions: Information, selection and marginalization. In *IPMI*, pages 435–446, 2009. 1

[26] M. Toews, W. M. Wells III, D. L. Collins, and T. Arbel. Feature-based morphometry: Discovering group-related anatomical patterns. *NeuroImage*, 49(3):2318–2327, 2010. 1, 4

[27] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affinely invariant neighbourhoods. *IJCV*, 59(1):61–85, 2004. 1, 2

[28] A. P. Witkin. Scale-space filtering. In *JCAI*, pages 1019–1021, 1983. 1, 2

[29] R. A. Young. The gaussian derivative model for spatial vision: I. retinal mechanisms. *Spatial Vision*, 2:273–293, 1987. 2