INVARIANT AND PERCEPTUALLY CONSISTENT TEXTURE MAPPING FOR CONTENT-BASED IMAGE RETRIEVAL

Huizhong Long, Chee Wee Tan, and Wee Kheng Leow

School of Computing, National University of Singapore,
3 Science Drive 2, Singapore 117543, Singapore

ABSTRACT

Texture is an important visual feature for content-based image retrieval. An ideal content-based retrieval system should compare images in its database with the query in a manner that is consistent with human's perception of visual similarity. Moreover, texture matching should be invariant to texture scale and orientation because the same texture can appear in the images in varying scales and orientations. In practice, however, texture similarity computed using computational texture features is not necessarily consistent with human's perception. This paper presents a method of mapping texture features into a texture space that is scale and orientation invariant, and at the same time, consistent with human's perception. Test results show that this method achieves a better retrieval performance than methods that are not invariant and not perceptually consistent.

1. INTRODUCTION

Texture is an important visual feature for content-based image retrieval. An ideal content-based image retrieval system should compare images with the user-specified query in a manner that is consistent with human's perception. Moreover, texture matching should be invariant to texture scale and orientation because the same texture can appear in the images in varying scales and orientations.

Computational texture features commonly used for image retrieval include the statistical features of Tamura et al. [13] used in IBM's QBIC system, DFT in [12] and the Wold model [6], MRSAR in Wold model, Gabor features in [14], NeTra [9], and *invariant texture space* [5]. Among these models, [5, 12, 14] are scale and orientation invariant. However, our studies [8] show that texture similarity computed based on these methods are not necessarily consistent with human's perception. On the other hand, Santini and Jain's *fuzzy features contrast model* (FFC) [11] has the potential of accounting for various peculiarities of human's perceptual similarity.

This project is funded by NUS ARF R-252-000-049-112.

This paper presents a method of constructing a perceptual texture space that is consistent with human's perception, and a method of mapping computational texture features to the scale- and orientation-invariant perceptual texture space. Experimental results show that this method improves the precision and recall of texture image retrieval.

2. INVARIANT AND PERCEPTUAL MAPPING

The problem of mapping computational texture features into an invariant and perceptually consistent texture space can be viewed as a two-stage process. First, the features $\mathbf{t}_i$ of texture i in the *input space* $\mathcal{T}$ are mapped to a point $\mathbf{I}(\mathbf{t}_i)$ in an *invariant space* $\mathcal{I}$ that is invariant to scale and orientation. Then, $\mathbf{I}(\mathbf{t}_i)$ is mapped to a point $\mathbf{P}(\mathbf{I}(\mathbf{t}_i))$ in the *perceptual texture space* (PTS) $\mathcal{P}$. Functions $\mathbf{P}$ and $\mathbf{I}$ perform *nonlinear* mappings between *multidimensional* spaces such that the mean mapping error E is minimized:

$$E = \frac{1}{N} \sum_i [\, \mathbf{p}_i - \mathbf{P}(\mathbf{I}(\mathbf{t}_i))]^2 \tag{1}$$

where $\mathbf{p}_i$ is the position of texture i in PTS, and N is the number of textures. Note that it is not practically viable to perform this mapping in the reverse order $\mathbf{I}(\mathbf{P}(\mathbf{t}))$ because this would require the measurement of human's perception under various scales and orientations, which is too tedious to perform and too difficult to obtain reliable results. In the construction of the PTS, texture images are preprocessed to normalize texture intensity, contrast, scale, and orientation [7] so that texture similarity measured in PTS is not confounded by these factors.

2.1. Perceptual Texture Space

Several previous works attempted to measure human's perception of texture similarity by having human subjects rank textures along subjectively predefined properties such as directionality, coarseness, busyness, and etc. [13, 1] However, the subjects' interpretations of the meaning of these visual properties are expected to vary from one person to the other.

Therefore, it is uncertain whether the individual ranking results can be combined into group ranking results that represent the perception of a typical person. Ranking along predefined properties may also bias subjects' judgments towards these properties rather than other potentially salient features [10]. Moreover, recent psychological studies show that subjective texture properties cannot be reliably associated with the dimensions of a perceptual texture space [3]. Therefore, instead of ranking textures according to predefined properties, we constructed a PTS using an experimental procedure similar to those of [10, 3].

60 subjects were asked to perform a free sorting task on 60 texture images according to their perceptual similarity. These 60 images consisted of 50 textures selected from the Brodatz album, and 10 other natural textures (such as foliage and roof tiles) not contained in the Brodatz album. Other textures in the Brodatz album were not selected because they were either scaled versions of the 50 selected textures, or their contents (e.g., flowers) were much stronger than their texture patterns. It is very difficult for the subjects to avoid classifying the images based on their contents.

After obtaining the sorting results, pairwise distances between textures were computed using Donderi's information measurement [2]:

$$p_{ij} = \sqrt{H(t_i|t_j) + H(t_j|t_i)} \qquad (2)$$

where $H(t_i|t_j)$ is the *independent information* of t_i after knowing the grouping of t_j. Independent information is the information gained by classifying two stimuli into separate groups. Let G denote the set of textures used in the experiment and $G(t_i)$ denote the group in which t_i belongs. The independent information $H(t_i|t_j)$ is computed as follows:

$$H(t_i|t_j) = \begin{cases} 0 & \text{if } G(t_i) = G(t_j) \\ \log_2 |G| - \log_2 |G(t_i)| & \text{if } G(t_i) \neq G(t_j) \end{cases} . \qquad (3)$$

Based on the pairwise distances, multidimensional scaling was performed to derive perceptual spaces with different number of dimensions. Our study [8] showed that both 3D and 4D spaces correlate well with existing perceptual spaces of Rao and Lohse [10] and of Heaps and Handel [3]. Moreover, the 4D space fits the psychological data better than does the 3D space. Therefore, the 4D space is used in our invariant and perceptual mapping method.

2.2. Invariant Mapping Network

Determination of $\mathbf{I}$ and $\mathbf{P}$ can be a difficult problem. It is not possible to determine $\mathbf{I}$ and $\mathbf{P}$ separately because the optimal coordinates $\mathbf{I}(\mathbf{t})$ of the textures in the invariant space $\mathcal{I}$ are unknown and must be deduced by the optimization process. Our method optimizes both $\mathbf{I}$ and $\mathbf{P}$ at the same time using an invariant neural network (Fig. 1).

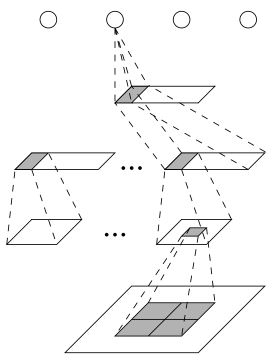

Fig. 1. Neural network for invariant and perceptually consistent mapping. The first three layers perform invariant mapping while the last three perform perceptual mapping.

The input layer of the neural network takes Gabor features of a texture as inputs. The Gabor features consist of 8 spatial frequency channels and 6 orientation channels. The second layer contains several feature maps. Each neuron in a feature map receives a local receptive field of inputs and the neurons in the same map have the same weight structure. That is, the second layer implements several translation-invariant maps, each is trained to detect a specific kind of pattern in the 2D Gabor inputs. This design is similar to the *convolutional network* [4] and is used to capture scale and orientation variations of textures which manifest as translation variations in its 2D Gabor pattern [5].

The third layer consists of a set of analyzers for each feature map. Each analyzer combines the results of all the neurons in the corresponding feature map. Therefore, the analyzers can detect a feature regardless of its position in the 2D Gabor inputs, i.e., they are scale and orientation invariant. The first three layers of the invariant network therefore perform the invariant mapping $\mathbf{I} : \mathcal{T} \rightarrow \mathcal{I}$ from input space $\mathcal{T}$ to an invariant space $\mathcal{I}$.

The last two layers are fully connected to their corresponding inputs. They perform nonlinear regressions between the activation values of the analyzers and the four output neurons, one for each dimension of PTS. That is, they perform the perceptual mapping $\mathbf{P} : \mathcal{I} \rightarrow \mathcal{P}$ from invariant space $\mathcal{I}$ to perceptual space $\mathcal{P}$.

The invariant network is trained using error back-propagation to optimize the functions $\mathbf{I}$ and $\mathbf{P}$ simultaneously by minimizing the mean mapping error (Eq. 1). After experimenting with various network parameters, it was found that the optimal configuration of the network consists of 3×3 local receptive fields and 4 sets of analyzers and feature maps.

3. TESTING RESULTS

To fully assess the network's mapping capability, performance under five different test conditions were evaluated:

Table 1. Mean squared errors of testing under various conditions.

	I_n	T_n	I_v	T_v	R
SVM	0.0052	0.216	0.244	0.245	0.240
NN	0.0099	0.238	0.0074	0.148	0.016

1. I_n: new instance; normalized scale, orientation
   9 samples were cropped from each of the 60 texture images. For each texture, 7 samples were used for training, 1 for validation, and 1 for testing.

2. T_n: new type; normalized scale, orientation
   All samples of 48 (80%) of the textures were used for training. All samples of 6 (10%) of the textures were used for validation, and the rest for testing.

3. I_v: new instance; various scales, orientations
   56 variations, with 7 scales and 8 orientations, were produced from each of the 60 textures. 80% of these variations were randomly selected for training, 10% for validation, and 10% for testing.

4. T_v: new type; various scales, orientations
   Similar to condition 2 except that samples with various scales and orientations were used.

5. R: random
   Of all the scale and orientation variations of the 60 textures, 80% were randomly selected for training, 10% for validation, and 10% for testing.

Ten-fold cross-validation tests were performed for all 5 conditions. Conditions 2 and 4 are the most stringent because their testing samples are perceptually different from their training samples. Their testing errors are expected to be larger than those of the other conditions. The last condition mimics a practical situation in which random samples at random scales and orientations are available for training.

To obtain the baseline results, Support Vector Machine (SVM) regression was applied to the five conditions as well. For SVM regression, Gaussian kernels were chosen because our tests showed that they perform better than polynomial and sigmoid kernels. Various parameter values of Gaussian kernels were tested to obtain the best performance.

Table 1 summarizes the testing results. Under the condition of normalized scale and orientation (conditions 1 and 2), the invariant network achieves a performance comparable to that of SVM regression. This is expected because both these methods are good at nonlinear regression. Also as expected, testing errors for condition 2 (new type) are larger than those for condition 1 (new instance).

Under conditions 3, 4, and 5, SVM regression does not perform as well as the invariant network because the network is designed to capture variations of scale and orientation while the SVM is not. SVM's testing errors for these three conditions are roughly the same as that for condition 2 (new type, normalized scale and orientation). These results indicate that SVM treats textures with different scales and orientations as new texture types. On the other hand, the invariant network performs texture mapping in a manner that is invariant to scale and orientation. The invariant network performs better in conditions 3 and 4 than in conditions 1 and 2 because there are more training samples in the former conditions than in the latter.

4. APPLICATION EXAMPLE

To further assess the invariant network's performance, texture image retrieval tests were performed. In a similar test conducted by Ma and Manjunath [9], human subjects were asked to classify similar Brodatz textures into the same group. During retrieval, textures in the same group as the query are considered as relevant. In our tests, texture groupings were derived from k-means clustering of the textures in PTS. Out of the 60 textures in PTS, 20 groups were derived, giving an average of 3 textures per group.

Three types of retrieval tests were performed to assess different aspects of the invariant network:

1. Assessment of the benefit of perceptual mapping.
   Texture images with normalized scale and orientation were used. Invariant network's performance was compared with those of SVM regression and *fuzzy features contrast model* (FFC) of Santini and Jain [11]. Various parameter values of FFC were tested to obtain the best performance. In addition, scaled Euclidean distance was also tested because it was compared with FFC in [11].

2. Assessment of the benefit of invariant mapping.
   Textures with various scales and orientations were used. Invariant network's performance was compared with that of SVM regression.

3. Assessment of the combined benefit of invariant and perceptual mapping.
   Textures with various scales and orientations were used. Invariant network's performance was compared with those of SVM regression, FFC, and scaled Euclidean.

In these tests, invariant network and SVM regression trained under the random condition in Section 2.2 were used since random condition is similar to practical application condition. For all the tests, 30 textures were randomly selected as the query textures, and the average normalized precision and average normalized recall are reported in this paper.

Table 2 summarizes the test results for retrieval condition 1. Both invariant network and SVM regression have

Table 2. Results of retrieving textures with normalized scale and orientation.

	norm. precision	norm. recall
invariant network	0.589	0.939
SVM regression	0.622	0.946
FFC	0.291	0.803
scaled Euclidean	0.143	0.781

Table 3. Results of retrieving textures with various scales and orientations.

	norm. precision	norm. recall
invariant network	0.639	0.959
SVM regression	0.324	0.833
FFC	0.269	0.650
scaled Euclidean	0.079	0.582

been trained to map texture features to PTS. Therefore, their performances are better than those of FFC and scaled Euclidean. FFC has a better performance than does scaled Euclidean distance, indicating that it can indeed capture some aspects of human perception.

Table 3 summarizes the test results for retrieval conditions 2 and 3. Invariant network has the best performance because it can perform both invariant and perceptual mappings well. Although SVM regression has been trained to perform perceptual mapping, it does not capture the variations of texture scale and orientation. Therefore, it does not perform as well as the invariant network. FFC and scaled Euclidean distance have lower precision and recall than do invariant network and SVM because they do not perform invariant and perceptual mappings well.

5. CONCLUSION

This paper presented a neural network that is trained to optimize both invariant and perceptual mappings. The first three layers of the network map the Gabor features of a texture to an invariant space, and the last three layers map the texture to the perceptual space, which is constructed using data obtained from psychological test of human's perception. Extensive tests were conducted to assess the performance of the invariant network. Test results show that the invariant network can perform invariant and perceptual mappings accurately, and invariant and perceptual mappings indeed improve the performance of texture image retrieval.

6. REFERENCES

[1] N. Abbadeni, D. Ziou, and S. Wang. Computational measures corresponding to perceptual textural features. In *Proc. ICIP*, 2000.

[2] D. C. Donderi. Information measurement of distinctiveness and similarity. *Perception and Psychophysics*, 44(6):576–584, 1988.

[3] C. Heaps and S. Handel. Similarity and features of natural textures. *J. Expt. Psycho.: Human Perception and Performance*, 25(2):299–320, 1999.

[4] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, 1995.

[5] W. K. Leow and S. Y. Lai. Scale and orientation-invariant texture matching for image retrieval. In M. K. Pietikäinen, editor, *Texture Analysis in Machine Vision*. World Scientific, 2000.

[6] F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. PAMI*, 18(7):722–733, 1996.

[7] H.-Z. Long and W. K. Leow. Perceptual texture space for content-based image retrieval. In *Proc. Int. Conf. on Multimedia Modeling (MMM 2000)*, pages 167–180, 2000.

[8] H.-Z. Long and W. K. Leow. Perceptual texture space improves perceptual consistency of computational features. In *Proc. IJCAI*, 2001.

[9] W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. In *Proc. CVPR*, pages 1160–1169, 1996.

[10] A. R. Rao and G. L. Lohse. Identifying high level features of texture perception. *CVGIP: Graphical Models and Image Processing*, 55(3):218–233, 1993.

[11] S. Santini and R. Jain. Similarity measures. *IEEE Trans. PAMI*, 21(9):871–883, 1999.

[12] D.-G. Sim, H.-K. Kim, and D.-I. Oh. Translation, scale, and rotation invariant texture descriptor for texture-based image retrieval. In *Proc. ICIP*, 2000.

[13] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. SMC*, 8(6):460–47, 1978.

[14] A. Teuner, O. Pichler, J. E. Santos Conde, and B. J. Hosticka. Orientation- and scale-invariant recognition of textures in multi-object scenes. In *Proc. ICIP*, pages 174–177, 1997.