Petr BEREK, Michal PRÍLEPOK, Jan PLATOŠ, Václav SNÁŠEL

VSB – Technical University of Ostrava

Similarity detection of image using vector quantization and compression

Abstract. *In every day is a lot of new images and photos get into the internet. Problem of image similarity is up-to-date in image retrieval. There are a lot of methods for comparison of images. We use vector quantization and NCD method for look for similar images in collection that vector quantization prepares image files for NCD. In this paper we can show how to convert 2D image into 1D string using by vector quantization and how NCD method is used for image similarity detection.*

Streszczenie. *W artykule analizowano problem podobieństwa obrazu. Użyto metody kwantyzacji wektora i metody NCD. Pokazano jak konwertować obraz 2D w strumień 1D. (**Detekcja podobieństwa obrazu bazująca na kwantyzacji wektora i kompresji**)*

Keywords: image similarity, compression, NCD, vector quantization
Słowa kluczowe: podobieństwo obrazu, kompresja, kwantyzacja wektora

Introduction

A lot of method for images similarity exists. For comparison of images we need some properties from images. In next we compare these properties. As first properties we can use colour property [1, 2]. Basic method for comparison by colour is pixel-by-pixel method. But this method is useful for closeness images with small divergence.[3] Next method can be comparison by colour histogram [4]. Statistically it means the join probability of intensities of three colour channels. But we need compare to two histograms. We can mention L2-related metric for comparison of two histograms [5, 6]. In next we can created colour sets and use quantization or we can use moment colour approach [7].

We can use textures also. Textures are importance and usefulness in pattern recognition and computer vision. It is property of surfaces as cloud, trees, bricks, fabric and hair. Textures can give important information about structural of surface. There are two approaches for textures comparison. [1] First is co-occurrence matrix. Base of this approach is co-occurrence matrix based on orientation and distance between pixels and then extracts meaningful statistics from matrix as the texture representation.[7] Second approach is based on psychological studies in human visual perception of texture. In psychological studies these properties were found: coarseness, contrast, directionality, linelikeness, regularity, and roughness. [8]

Different method how we can compare two images is compare their tags. Tags can be created manually or automatically. Tags give some text information about content of image – person, clouds, party, etc. We give these tags and we look for similar images. There can be question how can be tags accuracy? It is depend on people who describe image or automatically system.

NCD

The Normalized Compression Distance (NCD) is based on Kolmogorov complexity. It makes use of standard compressors in order to approximate Kolmogorov complexity. The NCD has been used for text retrieval [10], text clustering, plagiarism detection [4], music clustering [6, 10], music style modeling [11], automatic construction of the phylogeny tree based on whole mitochondrial genomes [12], the automatic construction of language trees [13, 14], and the automatic evaluation of machine translations [15].
The NCD is a mathematical way for measuring the similarity of objects. Measuring of similarity is realized by the help of compression where repeating parts are suppressed by compression. NCD may be used for comparison of different objects, such as images, music, texts or gene sequences. NCD has requirements
to compressor. The compressor meets the condition

$$C(x) = C(xx)$$

within logarithmic bounds [17]. We may use NCD for detection of plagiarism and visual data extraction[18, 19].
The resulting rate of probability distance is calculated by the following formula:

$$NCD = \frac{C(xy) - \min(C(x), C(y)}{\max(C(x), C(y)}$$

where:
- C(x) ist he length of compression of x.
- C(xy) is the length of compression concatenation of x and y.
- min(x; y) is the minimum of values x and y.
- max(x; y) is the maximum of values x and y.

The NCD value is in the interval 0 < NCD(x; y) < 1+epsilon. If NCD(x; y) = 0, then files x and y are equal. They have the highest difference when the result value of NCD(x; y) = 1+epsilon. The constant epsilon describes the inefficiency of the used compressor.
The NCD is not a metric. It is an approximation of the NID. The computation of the NCD is very efficient because we do not need to create the output itself. We compute only the size of the output. A study of the efficient implementation of the compression algorithms may be found in [20].

We can use for image similarity by NCD. But we need translate some properties from image into text. With above in text we can use text description image by tag. But we need properties from image. Is possible to create 1D string from 2D image? This process is called as linearization. But can 1D string represent 2D image? This approach is described in [3].

Experiment

Experiment setup

We use vector quantization [21] method for linearization 2D image into 1D string. We divided image into small pieces created by 8x8 pixels. We used quaternion for colour representation [22] of each pixel. So, we created blocks of 8x8 quaternions and codebook was created by these blocks [23]. We looked for coefficient C for relation between codebook compression and image quality after vector quantization. In experiment on collection described below we founded C as 2/3. Algorithm for block clustering is showed on Fig. 1.

```
CB is codebook
|Cs|=size of CB //count of blocks in CB
|Cn|=size of CB //count of blocks in CB
while(Cn > Cs x C)
{
        Find minimal distance between all pair of blocks
        B and NB are blocks with minimal distance between them
        create new block AB as average NB and B and save AB into Cl
        delete NB and B from CB
        |Cn|=|Cn|-1
}
```

Fig. 1 Algorithm of the Vector Quantization

Collection

We used collection from the internet [24]. There are over 850 images. In collection, there are images in resolution 786x576 pixels but for our experiment we changed resolution at 314x235 pixels. Images are divided into a lot of directories by groups (Animals, Flowers, Foliage, Textures, Fruits, Landscapes, Winter, Man, Mode, Shadows).

Saving image for using NCD

We used LZW compress method. This method needs string data, we needed to save image as a string. In first we tried to save only indexes from codebook. For each original block we looked for the nearest block in codebook after vector quantization and save index from codebook. Results of this approach were very poor. In next we saved all blocks in codebook and indexes. Each quaternion in each block we saved as string we saved vector part (i.e. rgb parts) as string. We gave text file for compression.

Result

Table 1 shows values given by NCD method. We created result in directory stated inside table. We demonstrate similarity detection of images only on images from directories in table 1. Results created from these directories were the best and results are most presentable. On figures below we show example of similar images. First image is origin image too. Search algorithm founds same image as the nearest image of original image. Each example has five images. Number below images is number get by NCD method.

Table 1: Results of the Similarity detection

Directory	Min value	Max value	Average value
Animals	0,9263	1,1028	0,968
Flowers	0,8376	1,0499	0,8787
Foliage	0,9179	1,0896	0,9719
Textures	0,9151	1,1021	1,0069

Conclusion

In this paper we showed method for image similarity detection based on vector quantization and NCD method. In example we showed this method is functional and this method has practical use. In next paper we want to use another larger collection of image and we want to try functionality of this method for larger collection. We want to try better solution of saving codebook after vector quantization and we want to create solution protect from image rotation and noise inside image.

Acknowledgment

This work was partially supported by the Grant Agency of the Czech Republic under grant no. P202/11/P142, and by the Grant of SGS No. SP2013/70, VSB - Technical University of Ostrava, Czech Republic, and was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

REFERENCES

[1] Yong Rui, Thomas S. Huang, Image Retrieval: Current Techniques, Promising Directions And Open Issues, Journal of Visual Communication and Image Representation (1999), No. 5, pages 39-62

[2] C. S. McCamy, H. Marcus, and J. G. Davidson, A color-rendition chart, Journal of Applied Photographic Engineering 2(3), 1976.

[3] Jonathan Mortensen, Jia Jie Wu, Jacob Furst, John Rogers, Effect of Image Linearization on Normalized Compression Distance

[4] M. Swain and D. Ballard, Color indexing, International Journal of Computer Vision 7(1), 1991.

[5] M. Ioka, A Method of Defining the Similarity of Images on the Basis of Color Information, Technical Report RT-0030, IBM Research, Tokyo Research Laboratory, Nov. 1989.

[6] W. Niblack, R. Barber, and et al., The QBIC project: Querying images by content using color, texture and shape, in Proc. SPIE Storage and Retrieval for Image and Video Databases, Feb. 1994.

[7] M. Stricker and M. Orengo, Similarity of color images, in Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995.

[8] R. M. Haralick, K. Shanmugam, and I. Dinstein, Texture features for image classification, IEEE Trans. On Sys. Man. and Cyb. SMC-3(6), 1973.

[9] H. Tamura, S. Mori, and T. Yamawaki, Texture features corresponding to visual perception, IEEE Trans. On Sys., Man. and Cyb. SMC-8(6), 1978.

[10] A. Granados. Analysis and study on text representation to improve the accuracy of the normalized compression distance. AI Commun., 25(4):381-384, 2012.

[11] S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano. Using machine-learning methods for musical style modeling. Computer, 36(10):73-80, 2003. cited By (since 1996)25.

[12] M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An informationbased sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics, 17(2):149-154, 2001. cited By (since 1996)274.

[13] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. Physical Review Letters, 88(4):487021-487024, 2002. cited By (since 1996)145.

[14] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi. The similarity metric. IEEE Transactions on Information Theory, 50(12):3250{-264, 2004.

[15] J. J. Vayrynen, T. Tapiovaara, K. Kettunen, and M. Dobrinkat. Normalized compression distance as an automatic MT evaluation metric. In Proceedings of MT 25years on, 21-22 Nov 2009 Craneld, UK, to appear

[17] D. Sculley and C. Brodley. Compression and machine learning: A new perspectiveon feature space vectors. pages 332-341, 2006. cited By (since 1996)17.

[18] P. M. B. Vit‼ nyi. Universal similarity. CoRR, abs/cs/0504089, 2005.

[19] R. Cilibrasi and P. M. B. Vitanyi. Clustering by compression. IEEE Transactionson Information Theory, 51(4):1523-1545, 2005.

[20] J. Walder, M. Kratky, R. Baca, J. Platos, and V. Snasel. Fast decoding algorithms for variable-lengths codes. Inf. Sci.,183(1):66-91, 2012.

[21] A. Gersho and R. M. Gray. Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA, USA,

[22] J. Angulo. Computational color imaging. chapter Structure Tensor of Colour Quaternion Image Representations for Invariant Feature Extraction, pages 91-100. Springer-Verlag, Berlin, Heidelberg, 2009.

[23] C.-Y. C. Tzu-Chuen Lu. A Survey of VQ Codebook Generation. Journal of Information Hiding and Multimedia Signal Processing, 2010.

[24] http://tabby.vision.mcgill.ca/html/welcome.html. [Online; accessed 11.4.2013]

Fig. 2 Similar images from Animals

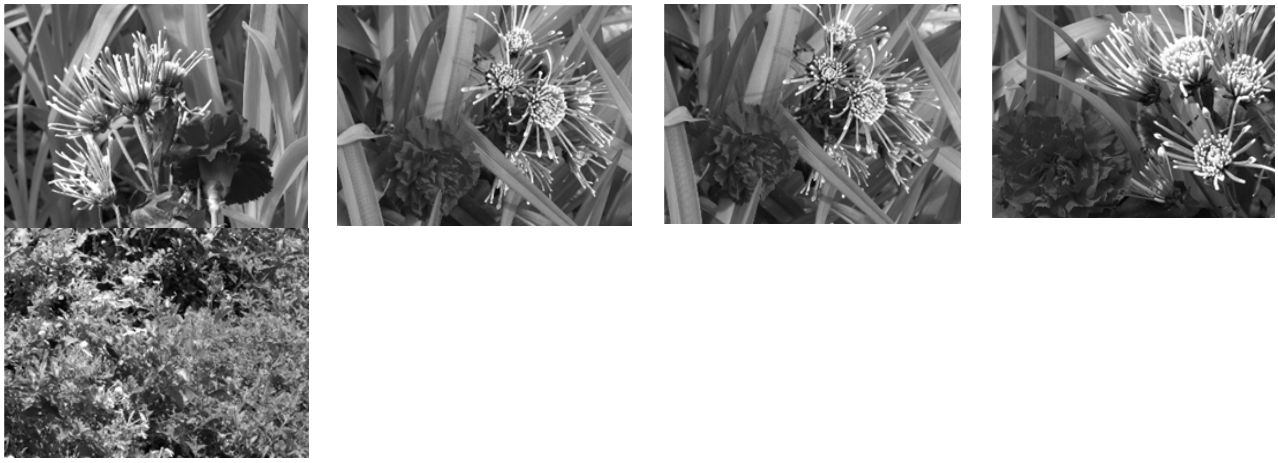

Fig. 3 Similar images from Flowers

Fig. 4 Similar images from Foliage

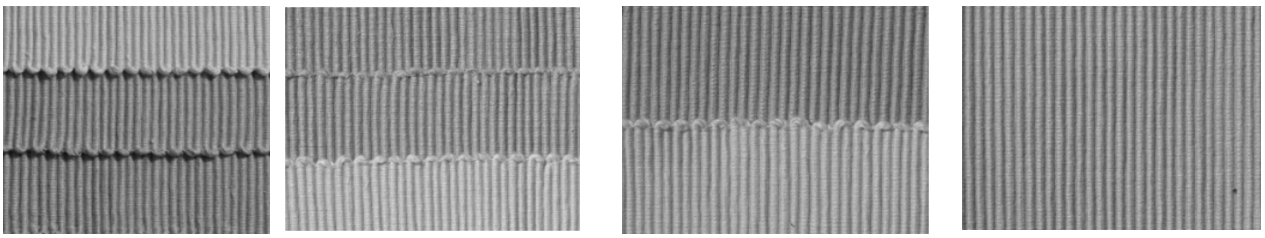

Fig. 5 Similar images fro Textures

***Authors**: Petr Berek,VSB-Technical University of Ostrava, 17 listopadu 15,70833 Ostrava Poruba, E-mail:petr.berek@vsb.cz; Michal Prilepok,VSB-Technical University of Ostrava, 17 listopadu 15,70833 Ostrava Poruba, E-mail:michal.prilepok@vsb.cz; Jan Platos, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:jan.platos@vsb.cz; Vaclav Snasel, VSB-Technical University of Ostrava, 17. listopadu 15,70833 Ostrava Poruba, E-mail:vaclav.snasel@vsb.cz;*