Image Processing

Pinar Duygulu*, Adrian Popescu**

* Bilkent University, Turkey

**CEA, LIST, LVIC France

Contacts: duygulu@cs.bilkent.edu.tr, adrian.popescu@cea.fr

MUCKE Project, Deliverable 2.2

08/04/2014

Contents

Abstract

MUCKE aims to mine a large volume of images, to structure them conceptually and to use this conceptual structuring in order to improve large-scale image retrieval. The last decade witnessed important progress concerning low-level image representations. However, there are a number problems which need to be solved in order to unleash the full potential of image mining in applications. The central problem with low-level representations is the mismatch between them and the human interpretation of image content. This problem can be instantiated, for instance, by the incapability of existing descriptors to capture spatial relationships between the concepts represented or by their incapability to convey an explanation of why two images are similar in a content-based image retrieval framework. We start by assessing existing local descriptors for image classification and by proposing to use co-occurrence matrices to better capture spatial relationships in images. The main focus in MUCKE is on cleaning large scale Web image corpora and on proposing image representations which are closer to the human interpretation of images. Consequently, we introduce methods which tackle these two problems and compare results to state of the art methods.

Note: some aspects of this deliverable are withheld at this time as they are pending review. Please contact the authors for a preview.

1 INTRODUCTION

The purpose of this report is to describe work on different aspects of image processing techniques utilised and introduced in MUCKE for handling multimedia information streams shared on social media. The readers of this report are highly likely to be computer literate individuals, with at least a college degree.

MUCKE aims the mining of large number of web images. This requires to attack several problems including parsimonious image description, large scale concept detection, detector generalisation across different datasets, etc. In this report, we present our current work in dealing with these problems.

We assess the use of different local image descriptors, as well as their combination with other properties, such as colour. The first part of the report focuses on these low level visual features for representing the images. First, some of the low-level features which are widely used in the literature for object and scene categorisation will be briefly described. Then, Vocabulary Trees which we utilised for scalable scene classification will be introduced.

One important drawback of the low-level features is their incapability in capturing spatial relationships. We revised co-occurrence matrices in order to capture local arrangements of low-level features, and proposed a new descriptor for efficient representation.

The second part of the report is dedicated to two novel approaches developed within the project for high level image representation and automatic concept learning from noisy web data.

In order to handle the insufficiency of the predefined set of detectors, to capture the entire set of concepts, we developed a systems to learn visual concepts from user tags (without requiring manual labels) for automatic concept detection and annotation purposes. First, Web images with noisy tags are used for collecting an initial pool of data for each concept. These images are then pruned from outliers and organised into clusters. We proposed Rectifying Self Organising Maps (RSOM) as a novel clustering and outlier detection method. Clusters obtained by RSOM are then used to model different properties of concepts.

The mismatch between low-level image representations and their understanding by humans is a well known problem in multimedia mining. We contribute to the reduction of this mismatch by introducing a high-level image representation of image content which makes use of large-scale image classification in order to identify the concepts represented in the image. This representation builds on recent advances in computer vision, brought by the introduction introduction of large neural networks. Image retrieval experiments show that the new representation compares favorably to state of the art representations and, equally important, is scalable.

These representations are relevant for two important areas of MUCKE: it will facilitate modality integration (Tasks 2.3 and 4.2) and can be used as an efficient alternative to existing low-level features in the MUCKE prototype (Task 5.3).

2 STATE-OF-THE-ART VISUAL DESCRIPTORS

In this section, we describe some of the low level visual features used in our studies. We will focus on three state-of-the-art features, namely Scale Invariant Feature Transform (SIFT), Histograms of Oriented Gradients (HOG) and GIST which are widely used in the literature. Other features utilised include colour histograms and Local Binary Patterns[8].

2.1 SCALE INVARIANT FEATURE TRANSFORM (SIFT)

Scale Invariant Feature Transform (SIFT) has been proposed by Lowe [5] and used in wide range of areas such as object recognition, 3D modelling, image stitching, video tracking, etc. SIFT allows the key-points (interest points, salient points) detected in an image to have a representation invariant to translation, scaling and rotation.

Lowe uses Difference of Gaussians (DoG) function to determine key-points. DoG is applied to a series of smoothed and resampled images and maxima and minima of the results are used in determination of key points. Then, low responses are filtered from the set of candidate key-points. Orientation of a key-point is assigned based on the dominant orientation of gradients around the key-point. Key-points are described by the distribution of gradients for 4x4 subregions in 8 bins, resulting in 128 length feature vector.

SIFT descriptors are generally used with Bag of Words (BoW) model in computer vision [3]. To represent an image with BoW model an image is treated as a document. Features are quantised to generate a codebook, and images are represented by the histogram of words from the codebook.

2.2 HISTOGRAMS OF ORIENTED GRADIENTS (HOG)

Introduced by Dalal and Triggs in [1], Histogram of Gradients (HOG) is a popular feature descriptor that is used widely in computer vision domain. It captures gradient structures that are characteristic of local shape. HOG method finds gradient orientations on a dense grid of uniformly spaced cells on an image, and quantises gradients into histogram bins. Local shape information is well described by the distribution of gradients in different orientations.

2.3 GIST

GIST descriptor is first proposed in [9] for scene recognition. It is based on low dimensional representation of the scene that is called Spatial Envelope [9] They define the features that separates a scene from the rest. Those features that represent dominant spatial structure of a scene are naturalness, openness, roughness, expansion and ruggedness. A multidimensional space is created to find out which scenes share membership in semantic categories such as street and highways by projecting shared memberships are projected closed together. Success of GIST in scene recognition supports that modelling a holistic representation of the scene is informative enough about scenes semantic categories.

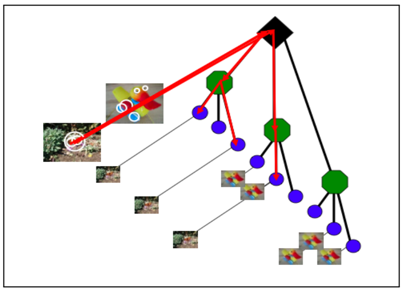

FIGURE 1: ILLUSTRATION OF VOCABULARY TREE [?]

3 VOCABULARY TREE FOR LARGE SCALE SCENE RECOGNITION

Vocabulary tree approach is proposed by Nister in [7] for object recognition. It is a recognition scheme that can efficiently scale on large number of objects. We exploit vocabulary tree approach for scene classification and landmark recognition.

Different than classical BoW approach vocabulary tree approach uses hierarchical k-means clustering and the tree directly defines the quantisation. To describe the images interest points are found and described. Nister uses Maximally Stable Regions [6] to find interest points and create descriptors of these regions with SIFT [4]. After extracting descriptors of each image in the database Nister creates the tree. k defines the branching factor of the tree at each level in k-means. First, an initial k-means process is run on the database and first k clusters are defined, each cluster having the descriptors closest to their centroids. This process is recursively applied to each group of descriptors and this recursive approach defines quantisation cells by splitting each quantisation cell into k new parts. This process goes up to some pre-defined tree depth. At the end of the tree construction we have our visual vocabularies which are the centroids.

To represent each image descriptors are pushed down the tree and the path is tracked that the descriptors follow. By counting how many times each node of the tree is visited by the image's descriptors they create a feature vector. To represent an image, and a tree with branching factor k and depth L, we need k dot products at each level of the tree. Illustration of the approach is given in the Figure 1.

This approach is also efficient in terms of memory usage. To represent the tree we do not need to store the image descriptors. Because we only need the tree to push the query image's descriptors.

Relevance of a database image to the query image is defined by how similar the paths down the vocabulary tree are. Nister also defines an entropy weighting for the nodes of the tree based on how much each node in the tree is visited during the training and they use this weights while calculating the image feature vectors.

Retrieval efficiency of this work comes from the inverted file idea. For each node keeping track of the images whose descriptors passed through that node reduces search space drastically.

4 A NEW DESCRIPTOR BASED ON CO-OCCURRENCES

Gray-level co-occurrence matrices (GLCM) have been introduced by Haralick et. al. [2] for analysis of textures. Given two values i and j, and offsets δx and δy, a co-occurrence matrix C encodes the distribution of co-occurring values at that given offset [1].

$$C_{\delta x,\delta y}(i,j)=\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \text{ and } I(p+\delta x,q+\delta y)=j\\ 0, & \text{otherwise}\end{cases}$$

Here p and q are the spatial positions in the image I of size nxm and the offset $\delta x,\delta y$ depends on the direction θ and the distance d at which the matrix is computed. The value of the image originally referred to the grayscale value of the specified pixel, but could be anything as will be discussed below.

Generally, second order statistics are extracted from gray level co-occurrence matrices and used for analysis of grey-level textures. We utilise and extend the co-occurrence matrices for describing the spatial relationships between different type of features going beyond intensity values. Moreover, we propose a new descriptor to combine a set of co-occurrence matrices which is richer than the use of a few statistics.

We are inspired from the textons idea where a set of filter outputs are generated for a given texture using different filters, and then outputs of different filters are kept in a vector for each pixel. These vectors are quantised to generate a codebook, and textures are then represented as a distribution of words in the codebook.

In our case, we generated a set of output images by computing the co-occurrences with different offsets for a single input image. For example, if we only consider the intensity values, and 8-neighbourhood connectivity with 1-pixel apart, then eight different co-occurrence matrices, $C_k; k = 1\ldots 8$, of size NxN will be generated where N=256 is the number of intensity values used. We form a vector of size eight for each cell i,j in the co-occurrence matrices, and quantise these words to construct a codebook. Rather than representing an image with the distribution of words from the codebook, we use the full matrix of size NxN as the feature. Figure**??** depicts the overview of our approach.

We perform preliminary experiments to compare the new descriptor for analysis of textures on two benchmark datasets (see Tables 1 and 2). We also test its effectiveness for scene classification on a very simple setup using only intensity values (see Table 3). We refer to the results obtained through the use of original statistical features extracted from co-occurrence matrices as `Cooc_Statistics`. We generate a codebook from the set of co-occurrence matrices. We refer to the matrix of size the co-occurrence matrices, but each cell corresponding to a word from the codebook as `CoocSet`. When we use the this matrix as it is as an NxN feature vector we call the method as `CooSet_Full`, whereas if we look at the distribution of the words in a Bag of Words manner we call the method as

[1] http://en.wikipedia.org/wiki/Co-occurrence_matrix

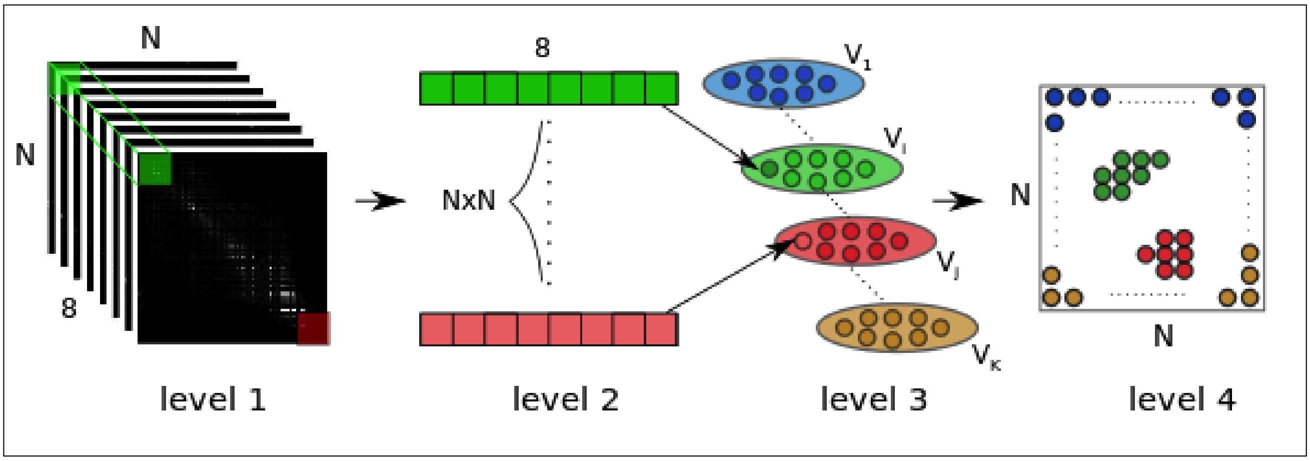

FIGURE 2: PROCESS OF PRODUCING N BY N HISTOGRAM MATRIX FROM GREY-LEVEL CO-OCCURRENCE MATRICES WITH K NUMBER OF WORDS. EACH PIXEL IS ALIGNED AT LEVEL 1 AND PRODUCES AN NXNXM MATRIX AS IN LEVEL 2, WHERE M IS THE NUMBER OF OFFSETS USED. THEN K NUMBER OF VISUAL WORDS ARE FOUND VIA CLUSTERING AT LEVEL 3. AT LEVEL 4 A HISTOGRAM MATRIX IS SHOWN WITH WORDS FOUND AT LEVEL 3.

Data	Cooc_Statistics	CoocSet_BoV	CoocSet_Full
D1	0.76	0.99	1.00
D2	0.59	0.78	1.00
D3	0.86	0.85	1.00
D4	0.72	0.98	1.00
D5	0.61	0.84	1.00
D6	0.88	0.89	1.00
D7	0.62	0.85	1.00
D8	1.00	1.00	1.00
D9	0.88	0.95	1.00
D10	0.63	0.84	1.00
Avg.	0.75	0.90	1.00

TABLE 1: MEAN AVERAGE PRECISION RESULTS ON BRODATZ DATASET.

Data	Cooc_Statistics	CoocSet_BoV	CoocSet_Full
S1	0.84	0.76	0.76
S2	0.92	0.80	0.78
S3	0.91	0.92	0.91
S4	0.72	0.69	0.73
S5	0.87	0.86	0.98
S6	0.77	0.80	0.81
S7	0.81	0.79	0.94
S8	0.91	0.80	0.95
S9	1.00	1.00	1.00
S10	0.66	0.79	0.90
Avg.	0.84	0.82	0.88

TABLE 2: MEAN AVERAGE PRECISION RESULTS ON CURET DATASET.

Data	Cooc_Statistics	CoocSet_BoV	CoocSet_Full
SC1	0.64	0.63	0.51
SC2	0.61	0.68	0.99
SC3	0.70	0.70	0.85
SC4	0.63	0.67	0.53
SC5	0.61	0.62	0.55
Avg.	0.64	0.66	0.69

TABLE 3: MEAN AVERAGE PRECISION RESULTS ON 15SCENE DATASET.

CoocSet_BoV. As can be seen from the results the proposed descriptor outperforms the statistical features, and use of the full matrix is better than using the distributions.

The proposed descriptor is not limited to intensity values, but could be extended for mid- and high-level features. Currently, we work on features such as SIFT, HOG and colour for scene classification.

5 CONCLUSIONS

In this report, we have summarised the work carried out in MUCKE for visual analysis of large number of images. Going beyond the use of state-of-the-art features, we have proposed new descriptors.

- We revisited co-occurrence matrices for constructing more powerful low level features that are also capable of capturing spatial relationships to some extent.
- We have proposed a method to learn mid- and high-level concepts from weakly labeled web images. The images obtained for a query concept are pruned from outliers, and grouped in order to capture different characteristics of the concepts with different models. We have shown that, mid-level attributes and scents as well as high-level objects and faces can be learned automatically without manual labels.
- We have introduced an innovative high-level image feature which is adapted for large-scale image retrieval. This feature allies compactness and efficiency and compares favourably with state of the art low-level descriptions.

  The results obtained in image mining are core component of future work in MUCKE. First they will be integrated with text representations in a multimedia fusion method. Second, fusion results will be exploited to compute multimedia concept similarities. Finally, image representations will be integrated in the MUCKE retrieval framework in order to rerank and diversify retrieval results.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[2] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610–621, Nov 1973.

[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[5] D. G. Lowe. Òdistinctive image features from scale-invariant keypointsÓ. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[7] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971Ð987, 2002.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.