Instant Bag-of-Words served on a Laptop

Jasper Uijlings
University of Amsterdam
Amsterdam, The Netherlands
jrr.uijlings@uva.nl

Ork de Rooij
University of Amsterdam
Amsterdam, The Netherlands
orooij@uva.nl

Daan Odijk
University of Amsterdam
Amsterdam, The Netherlands
d.odijk@uva.nl

Arnold Smeulders
University of Amsterdam
Amsterdam, The Netherlands
a.w.m.smeulders@uva.nl

Marcel Worring
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

ABSTRACT

This demo showcases our realtime implementation of concept classification using the Bag-of-Words method embedded within MediaTable, our interactive categorization tool for large multimedia collections. MediaTable allows the users to open images from disk or download these directly from the internet. Each image is then processed using the Bag-of-Words method, which computes classification scores for 20 distinct concepts classes on the fly. These are then seamlessly displayed in the interface.

1. INTRODUCTION

Over the last years the amount of available digital multimedia content has exploded. Social network sites Flickr or YouTube host respectively billions of images and many lifetimes of video. Many broadcasting companies are digitizing all media content in their archives. To make all this information accessible one needs to search within its content, where especially the visual search is computationally expensive.

In recent work [4] we have addressed the computational efficiency of the dominating framework in content-based image- and video retrieval: Bag-of-Words. With only algorithmic improvements (we do not use a GPU implementation), we accelerated the method by a factor 70 with a 3% loss in accuracy. This results in a realtime Bag-of-Words classification scheme on a standard desktop computer. In this demo we showcase this algorithm on a standard laptop with an interactive image categorization task in which interactive and automatic categorization are seamlessly integrated.

2. OUR DEMO

We have integrated our algorithm within MediaTable [1], our interactive multimedia categorization system. This system allows users to rapidly search and categorize collections of image or video material into user defined sets of related materials based on any kind of available metadata. MediaTable is typically used for search tasks, annotation tasks and categorization tasks. In this demo, we showcase MediaTable using 20 semantic concepts, which are extracted on the fly from the visual content itself.

We shall demonstrate the following demo. First, we allow users to select a set of images on hard disk, or by querying these from flickr. Next, these are shown in MediaTable, and the user can start categorizing them. Meanwhile, the algorithm starts to process these images, and the resulting classification scores are displayed in the interface as soon as they are available. Users can then inspect these scores immediately, or use them for further sorting and categorization of the displayed set of images. For more information see our pipeline in figure 1.

2.1 The Classes

We trained our method on the Pascal VOC 2010 trainval set. Hence we obtained classifiers for the following 20 classes:

aeroplane
bicycle
bird
boat
bottle
bus
car
cat
chair
cow
diningtable
dog
horse
motorbike
person
potted-plant
sheep
sofa
train
tv/monitor

3. CLASSIFICATION METHOD

We briefly describe the components that we use in our Bag-of-Words framework. For details we refer the reader to [4].

We densely sample SURF descriptors from the image which can be seen as a fast alternative to SIFT. We have accelerated the extraction of dense SURF features in two ways: 1) We sum haar responses over regions by using matrix multiplications, enabling the use of highly optimized matrix multiplication libraries. 2) We reuse measurements of the subregions of SURF. The Matlab code for calculating dense SURF (and also dense SIFT) has been made public[1].

As a visual vocabulary we do not use the standard k-means vocabulary with nearest neighbour assignment, but use a Random Forest as proposed by [3]. Hence visual words are assigned by multiple binary decision trees. We found that by first performing Principal Component Analysis on the descriptors we maintain a high accuracy [4]. We use

ICMR'11, April 17-20, Trento, Italy
ACM 978-1-4503-0336-1/11/04.

[1] www.science.uva.nl/~jrruijli

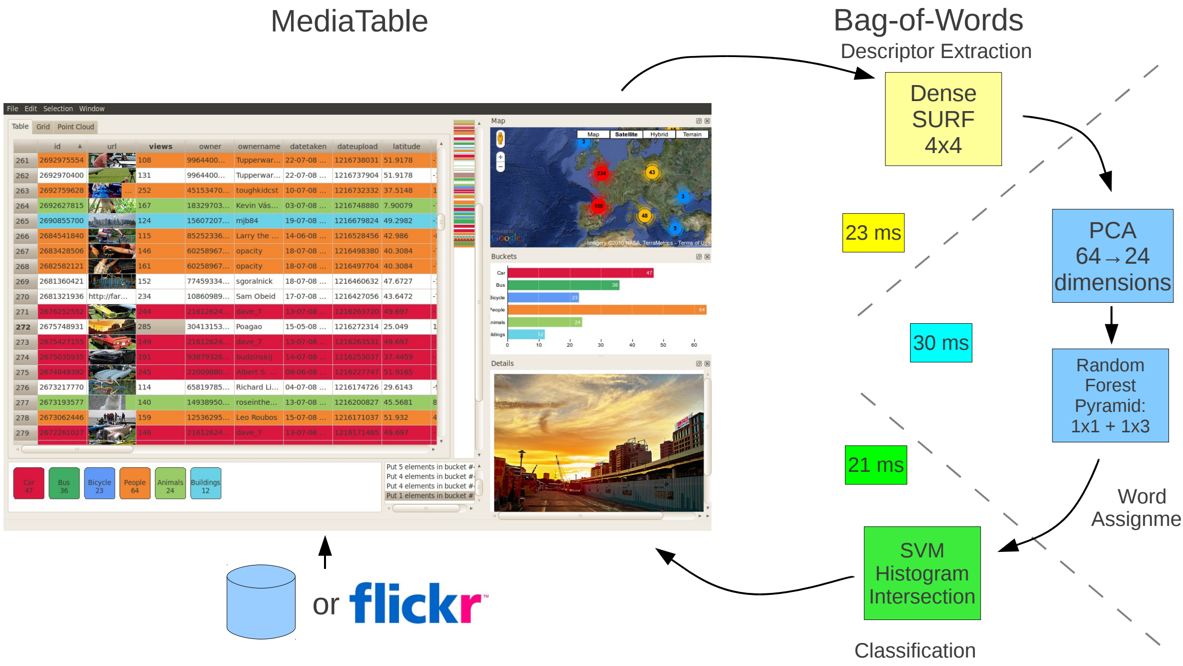

Figure 1: The MediaTable Bag-of-words pipeline. As soon as images are added they go through the bag-of-words pipeline. The results are then send back to MediaTable which updates the interface on the fly. On a single consumer laptop with an Intel Core 2 Duo T6400 2GHz processor this process takes 74 ms per image.

4 binary decision trees of depth 10, resulting in a visual vocabulary size of 4096.

We use the Spatial Pyramid with a subdivision in three horizontal regions.

For classification we use a Support Vector Machine with a Histogram Intersection kernel. We use the accelerated classification method of [2].

We showcase the demo within MediaTable on a single laptop with an Intel Core 2 Duo T6400 2GHz processor, not using the GPU. On this machine, the total Bag-of-Words framework takes 74 ms per image or can classify 14 images per second.

4. REFERENCES

[1] O. de Rooij, , M. Worring, and J. van Wijk. Mediatable: Interactive categorization of multimedia collections. *IEEE Computer Graphics and Applications*, 30(5):42–51, Sept. 2010.

[2] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:1632–1646, 2008.

[4] J. Uijlings, A. Smeulders, and R. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665 –681, 2010.