SAIVT-ADMRG @ MediaEval 2014 Social Event Detection

Simon Denman
SAIVT Laboratory
Queensland University of Technology
Brisbane, Australia
s.denman@qut.edu.au

David Dean
SAIVT Laboratory
Queensland University of Technology
Brisbane, Australia
d.dean@qut.edu.au

Clinton Fookes
SAIVT Laboratory
Queensland University of Technology
Brisbane, Australia
c.fookes@qut.edu.au

Sridha Sridharan
SAIVT Laboratory
Queensland University of Technology
Brisbane, Australia
s.sridharan@qut.edu.au

ABSTRACT

This paper outlines the approach taken by the Speech, Audio, Image and Video Technologies laboratory, and the Applied Data Mining Research Group (SAIVT-ADMRG) in the 2014 MediaEval Social Event Detection (SED) task. We participated in the event based clustering subtask (subtask 1), and focused on investigating the incorporation of image features as another source of data to aid clustering. In particular, we developed a descriptor based around the use of super-pixel segmentation, that allows a low dimensional feature that incorporates both colour and texture information to be extracted and used within the popular bag-of-visual-words (BoVW) approach.

1. INTRODUCTION

The Social Event Detection (SED) task at MediaEval 2014 [4] is concerned with the detection and retrieval of events from large multimedia collections. A key component of this social media is image and video data, which typically contains images or videos of the events taking place. However, in previous editions limited attention has been given to this data source. For instance in the 2013 evaluation, only two of the approaches sought to incorporate image features and in both cases they simply applied well established techniques. Motivated by this, we seek to investigate the use of visual features to aid social event detection and clustering.

A limitation of existing widely used approaches such as SIFT [2] is the high dimensionality (32 dimensions), which leads to increased memory demands, and the need for large codebooks when used in a BoVW framework. Furthermore, descriptors such as SIFT use greyscale images, discarding colour information, and although SIFT descriptors can be computed across multiple channels to incorporate colour, this further increases dimensionality. Motivated by this, we propose a new a low dimensional descriptor that incorporates both colour and texture information though the use of super-pixel segmentation. We combine this approach with an existing text processing system [5] and evaluate it on subtask 1 (event based clustering of the media collection). The remainder of this paper is structured as follows: Section 2 outlines the proposed approach; Section 3 presents and discusses our results; and Section 4 concludes the paper.

2. PROPOSED APPROACH

We aim to explore the use of image features for social event detection. We use the text processing based approach of [5] to combine meta-data (text data, time-stamp, and location information) and with visual features. We employ the BoVW approach for generating a visual descriptor. Our baseline approach uses the SIFT descriptor extracted in dense manner (with a bin size of 4 and a step size of 8) with K-means used to generate a codebook. A limitation with SIFT is its high dimensionality, necessitating a large dictionary and high memory requirements, and the fact that it ignores colour information. To alleviate this, we propose a new feature based on super-pixel segmentation. Super-pixel segmentation aims to segment an image into a set of related pixels, such that each super-pixel is formed by a set of connected and similar pixels (see Figure 1). We use the SLIC approach of [1] to extract super-pixels, and set the target super-pixel size to 20, to ensure that features are extracted from a similar size image patch as dense SIFT. From each resultant super-pixel, we extract a set of features to describe it's colour and texture. The colour component is the average colour of the super-pixel in LAB colour space divided by a normalisation factor, C. The role of C is to ensure that the colour and texture information contribute approximately equally to the feature vector, and is set empirically using the development set. The texture component is a HOG descriptor computed from all pixels in the super-pixel. We use an 8-bin histogram, and do not perform any normalisation prior to computing the HOG.

The resultant feature vector for each super-pixel can then be given as:

$$F = \{F_L, F_A, F_B, F_{HOG,0}, F_{HOG,1}, F_{HOG,2}, F_{HOG,3}, F_{HOG,4}, F_{HOG,5}, F_{HOG,6}, F_{HOG,7}\}, \quad (1)$$

where F_L, F_A and F_B are the LAB colour features; and $[F_{HOG,0}..F_{HOG,7}]$ are the 8 bins of the HOG histogram.

We utilise these features within the BoVW framework to

(a) (b)

Figure 1: An example of super-pixel segmentation using the SLIC algorithm. Note that larger super-pixels are shown here for visualisation purposes.

build an image descriptor. A codebook is trained (using K-means or Fisher Vectors [3]) using features extracted from several thousand images. Subsequent images are then encoded using this codebook to generate a descriptor that encapsulates the content of the images; and these descriptors are compared to one another using Euclidean distance.

Finally, text and visual features are combined in the following manner:

$$sim(d,p) = \beta_1 sim^{cosine}(d,p) + \beta_2 sim^{time}(d,p) + \beta_3 sim^{gps}(d,p) + \beta_4 sim^{image}(d,p), \quad (2)$$

where $sim^{cosine}(d,p)$, $sim^{time}(d,p)$ and $sim^{gps}(d,p)$ are the similarity of the text, timestamps and GPS locations as computed by [5]; $sim^{image}(d,p)$ is the similarity of the image features; and β_i are weight parameters used to combine the different data sources. These weight parameters are learnt from the training data to maximise clustering accuracy on the training set. Entries are then clustered using the constrained method of [5], which uses document ranking to choose a neighbourhood of best candidates from which the best match is chosen.

3. EVALUATION

3.1 Runs

Our five systems are as follows:

1. Metadata only: an implementation of [5].
2. Metadata + SIFT/K-means/1000: Meta-data combined with an image representation using SIFT features and a 1000 word K-means codebook.
3. Metadata + proposed super-pixel feature (SP)/ K-means/1000: Meta-data combined with an image representation using the proposed feature and a 1000 word K-means codebook.
4. Metadata + SP/K-means/125: As with system 3, except the dictionary is now of size 125.
5. Metadata + SP/FV/125: As with system 4, except Fisher Vector encoding [3] is used instead of K-means.

We use C++ and VLFeat [6] to encode images.

3.2 Results

Results for subtask 1 are shown in Table 1. We note that the incorporation of image data does lead to an improvement, albeit only a small one, over the baseline with systems 2-5 all outperforming the text only system (1). Of note is that system 4 outperforms that of 3, suggesting that the larger codebook used in 3 resulted in overfitting and thus a poorer representation. The use of Fisher Vectors [3] instead of K-means also leads to a small improvement, as can be seen by the improvement from systems 4 to 5. It should be noted that a Fisher Vector encoding could not be produced for the SIFT features, even with a much smaller dictionary size, due to the higher dimensionality of the feature and larger memory requirements of the training process.

Run	F1	NMI	Div. F1
1	0.7443	0.8993	0.7426
2	**0.7525**	**0.9018**	0.7508
3	0.7517	0.9017	0.75
4	0.7523	0.9018	0.7506
5	**0.7525**	**0.9018**	**0.7509**

Table 1: Results for the five runs for subtask 1. Refer to Section 3.1 run descriptions.

We observe that with the exception of system 5, the dense SIFT approach of system 2 outperforms systems using the proposed feature (3 and 4). However, the proposed approach has a much lower memory footprint than the SIFT descriptor (for instance dense SIFT features extracted from the training data require 254GB of storage, while using the proposed approach requires only 10GB), leading to significant improvements in computational efficiency when learning codebooks, and encoding features.

4. CONCLUSIONS AND FUTURE WORK

We have described our submission to the MediaEval 2014 SED task. Our approach uses a new feature representation for images, which we utilize with the popular bag-of-words framework. This has been shown to offer comparable performance to the SIFT descriptor, at much greater computational and memory efficiency. Future work will continue to investigate the proposed approach. Factors such as the normalisation of colour and HOG features, the number of orientation bins, and the size of the super-pixels will all be investigated. Furthermore, the method used to combine the visual data with the meta-data will be further investigated and refined to better utilise the visual information.

5. ACKNOWLEDGMENTS

We would like to thank Taufik Sutanto and Richi Nayak from the ADMRG at QUT for their assistance in completing this evaluation.

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[2] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157, 1999.

[3] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156. Springer, 2010.

[4] G. Petkos, S. Papadopoulos, V. Mezaris, and Y. Kompatsiaris. Social event detection at MediaEval 2014: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, 2014.

[5] T. Sutanto and R. Nayak. The ranking based constrained document clustering method and its application to social event detection. In *Database Systems for Advanced Applications*, pages 47–60. Springer, 2014.

[6] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.