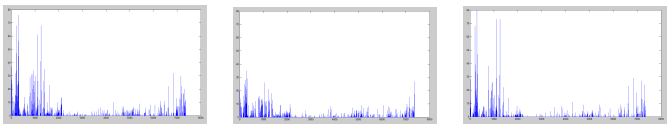
Push and Pull: Iterative grouping of media

Andrew Gilbert
www.andrewjohngilbert.co.uk
Richard Bowden
r.bowden@surrey.ac.uk

CVSSP
University of Surrey
Guildford, UK

While many techniques use the traditional approach of time consuming groundtruthing large amounts of data [1, 2], this is increasingly infeasible as dataset size and complexity increase. Instead we propose a solution that allows the user to select media that semantically belongs to the same class and use machine learning to "pull" this and other related content together. We use real data harvested from the internet and propose an approach capable of incrementally clustering similar material using the manual identification of a few true positive and false positive examples. In order to provide both scalability and incremental learning, the approach needs to be efficient. We combine two popular data mining tools developed for the text analysis domain to efficiently compute distances between high dimensional representations and dynamically augment the representation with new compound visual words to form an image signature. These tools are applied to selected true and false positive examples of the media and rules learnt, that are applied to the full corpus of material. The media is then formed into groups of same class media using a greedy clustering approach.

An image signature is constructed for each input sample; this is similar to a bag-of-words (BoW) histogram representation, and provides a compact, discrete representation of the input sample, as shown in Figure 1. In order to form the similarity between the image signatures, a data min-



(a) Shooting (b) Walking (c) Horse Riding

Figure 1: Image Signature class examples

ing tool called min-Hash is used and adapted. It is a randomised hashing approach, where the computation is proportional only to the number of input samples rather than the dimension of the vocabulary. To estimate the overlap of two image signatures, multiple independent min-Hash functions are used. The fraction of the min-Hash functions that assigns an identical value to the two sets gives an estimate of the similarity of the two image signatures. Figure 2 shows an example of the symbolisation of the image signature,

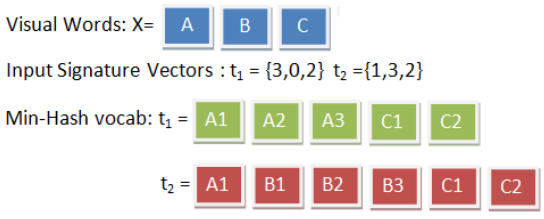


Figure 2: The approximation of a histogram as a visual word

The min-hash will return pairwise similarities between image signatures, therefore to efficiently cluster the classes, a greedy clustering approach is proposed. The aim is to form groups of image signatures based on the consistency of the min-Hash result. These groupings can then be visualised using Multidimensional scaling (MDS). It works by visualising the structure of the set of image signatures from the confusion matrix distances formed by the min-Hash. An example of the grouping and visualisation can be seen in Figure 4

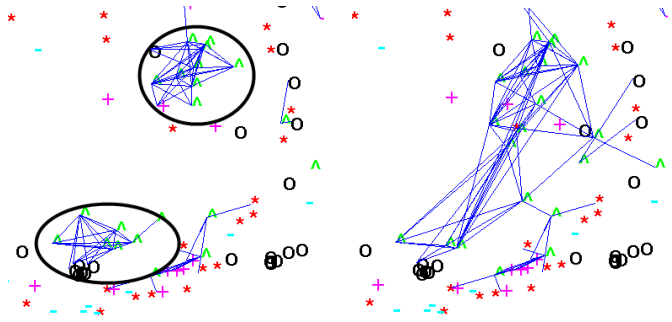
Initial visualisation of the min-Hash and the associated groups, will place many false postives within these groups. Therefore we propose to "push" false positive classifications apart and to "pull" false negatives closer together. To achieve this, a data mining tool is used, APriori association rule mining. It identifies elements of the image signatures that co-occur most frequently in the positive image signatures, with respect to the negative image signatures. This accentuates the elements common within the positive image signatures to pull them spatially closer, and this process is repeated.

The *YouTube* dataset was used to illustrate and evaluate the quality of the clustering and categorisation. Some examples of the dataset are shown in Figure 3.

(a) Cycling (b) Juggling (c) Basketball

Figure 3: *YouTube* dataset examples

The feature representation used by Gilbert *et al* [1] is employed, these are compound corner classifiers, that are based on a set of spatio-temporal Harris corner interest points. Each image signature for each video contains around 2000 elements, and the overall initial vocabulary of elements is 4458. Figure 4(a) shows a subset of the initial groupings for the class *Diving* from the *YouTube* dataset, each symbol represents a different class. It can be seen that there are a number of groups of true positive examples



(a) Initial greedy clustering result of the class Diving (b) Grouping of diving class after pulling together two separate groups

Figure 4: The lines indicate the grouping of the class diving from the *YouTube* dataset before and after pull groups together

but also many false positive. Overall for the *YouTube* dataset, there are initially 60.4% true positive groupings and 21.4% false positive groupings.

To improve this result the user can iteratively pull together groups of positive classifications. The aim is to pull together the two areas highlight by circles in Figure 4(a), performing the mining to identify common elements of the true positive image signatures and accentuating these to pull the true image signatures closer. Figure 4(b) shows the groupings after selecting all the videos within the two marked circled groups. In Figure 4(b), the two groups are more interlinked, this is reflected by the increased accuracy of correctly grouping *diving* examples by 10%. In addition, a number of the false positive links were removed as the true positive links have increased in strength.

[1] Andrew Gilbert, John Illingworth, and Richard Bowden. Action recognition usingmined hierarchical compound features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 883 – 897, 2011.

[2] I. Laptev and Pérez. "Retrieving Actions in Movies". *In Proc. of IEEE International Conference on Computer Vision (ICCV'07)*, 2007.