Balanced Clustering for Content-based Image Browsing

Tim Althoff, Adrian Ulges, Andreas Dengel
tim.althoff@dfki.de
German Research Center for Artificial Intelligence (DFKI)
and University of Kaiserslautern

Abstract: In recent years the explosive growth of digitally stored image and video data has raised the need for tools to search and organize visual data automatically by their content. Browsing environments, which structure image and video collections, are one solution to this problem. Therefore, image clustering techniques are needed that group semantically related images, are highly scalable, and produce balanced structures. We propose a simple and efficient strategy to enforce a more balanced clustering based on a hierarchical variant of the online k-means algorithm that favors small clusters over larger ones by adapting the prior probability of each cluster. We compare our method to standard hierarchical agglomerative techniques using multiple standard features and real-world datasets, showing that the proposed approach yields clusters of comparable qualitity while being substantially more balanced and scalable.[1]

GI-Topic: KI-BV (artificial intelligence - image understanding)

1 Introduction

As digital acquisition devices (still image and video cameras) are already wide-spread, it has become easy to acquire huge volumes of visual data. For example, users sharing their photos and videos on websites like Flickr and Facebook generate data in large amounts (Facebook registers around 2.5 billion photo uploads per month [Alt10]). Huge image and video collections are also becoming common due to commercial efforts like Google Street View or broadcasting networks [BSUB08]. This raises a need for efficient ways of organizing and structuring visual content such that users can browse and find the images and videos they are looking for.

One approach to solve this problem are content-based browsing environments Google Image Swirl or Navidgator [BSUB08], which offer an alternative to conventional approaches that rely on textual annotations of the data. Instead of presenting only a localized view of a few images, browsing environments also offer an overview of the entire database and conveniently allow the user to dynamically redefine his search. Internally, these content-based browsing systems rely on hierarchical clustering algorithms, which build a similarity-based structure upon the image database.

[1]This work was supported by the EU Safer Internet Programme, project FIVES (SIP-2008-TP-131801).

2 Our Approach

We use a hierarchical clustering algorithm that produces a nested tree of partitions called the *clustering tree*. Basically, there are three particular requirements that should be met by this clustering algorithm, namely quality, scalability and balancing [Alt10].

Quality The produced clusters should be of be a certain quality in a sense that images in a cluster are semantically related (e.g. showing the same scene or adhering to the same category). We measure quality by external cluster validity measures [CWK03, RH07].

Scalability The clustering algorithm needs to cluster large-scale image collections, using only a reasonable amount of resources. While many current systems have only been tested on datasets with 10k images [BSUB08, CBD04], clustering approaches should be able to process up to millions of images. Therefore, all approaches in $O(N^2)$ (particularly, hierarchical agglomerative clustering) – with N being the number of images – are not applicable. Our proposed method uses a k-means-based top-down approach in $O(N)$.

Balancing Clustering algorithms often produce unbalanced structures or trees that are difficult to navigate. Instead, regular structures help the user with browsing an image collection, orienting themselves in the dataset, and finding specific images more quickly. Thus, we demand that the clusters obtained are of approximately the same size, obtaining a balanced tree structure.

We choose a k-means-based approach [JMF99], which offers strong scalability and sufficient quality but unfortunately does not produce balanced clusters. To overcome this problem, Frequency Sensitive Competitive Learning algorithms (FSCL) has been suggested: while plain k-means uses the euclidean distance to decide to which cluster centroid a given data point should be assigned, FSCL approaches incorporate balancing by weighting the distance measure by the number of assignments to the respective cluster [AKCM90] (multiplicative bias). However, while these balancing approaches were demonstrated to work well on low-dimensional synthetic datasets, they were experienced as very unstable in high-dimensional feature spaces [Alt10]. To overcome this problem we combine two approaches in our proposed method `bo-k-means`, namely an online approach and frequency sensitive prior adaption.

Online Approach To encounter the instability of batch-based approaches, cluster assignments can be interleaved with cluster size estimation, resulting in an online approach that reestimates cluster sizes after each new assignment.

Frequency Sensitive Prior Adaption The key idea of our approach is to incorporate balancing by adapting the prior probability of each cluster to favor small clusters over larger ones. Whereas FSCL [BG04] assumed a uniform prior, it is now assumed to decrease for growing clusters. Mathematically, this leads to an additive bias instead of a multiplicative one, i.e. we *add* a penalty term depending on the cluster size to the actual distance of the sample to the cluster centroid. Sample x is assigned to the cluster k^* such that

$$k^* = \text{argmin}_k \ [(1-\beta) \cdot d(v_k, x) + \beta \cdot \alpha \cdot n_k] \quad 0 \leq \beta \leq 1$$

where $d(v_k, x)$ is the euclidean distance between the k-th cluster centroid v_k and the sam-

ple x, and n_k the size of the k-th cluster. β allows us to trade off true feature similarity against balancing constraints. We estimate a reasonable value for α from the given data such that for a value of $\beta = 0.5$ both distance-based similarity and cluster-size-based penalty are weighted equally on average (see [Alt10] for details).

3 Experiments & Results

We evaluated our method on several image datasets using various state-of-the-art features of content-based image retrieval [Alt10]. To measure cluster quality we used the external cluster validity measures Weighted Purity [CWK03] and V-Measure [RH07]. Balancing was evaluated using the Standard Deviation of Cluster Sizes (SDCS) [BG04] as well as the minimal and maximal cluster size.

Quality Our experiments showed that hierarchical k-means methods do not necessarily perform worse than agglomerative approaches, which represent the current state-of-the-art methods in content-based image clustering. In fact, they perform clearly superior for small numbers of clusters. However, the hierarchical agglomerative methods outperform the top-down approach on lower levels of the tree as they work bottom-up [Alt10].

Scalability While agglomerative methods cannot be reasonably applied to datasets beyond a few thousand images, our top-down approach proved to be substantially more scalable. For example, the Caltech-256 [GHP07] dataset (30k images) was clustered about 43 times faster by our proposed method compared to current agglomerative methods. We have also tested our approach on a large-scale dataset comprising over 700k keyframes from YouTube videos.

Balancing The proposed balancing method `bo-k-means` (and its hierarchical extension) yield highly balanced clusters while preserving or even improving cluster quality. Figure 1 illustrates this result on the Netclean dataset [HA06], a dataset of indoor picture series. A value for β of 0 implies no balancing at all while a value of 1 enforces all equal cluster sizes. Using a moderate balancing ($\beta = 0.4$), Weighted Purity increased by up to 33%. The proposed balancing approach was found to outperform FSCL methods in terms of stability. In addition, the strong regularization effect of balancing renders the clustering result almost initialization-independent, significantly alleviating the inherent problem of initialization-dependence of k-means-based methods.

Figure 2 shows example clusters of our proposed algorithm on the well-known Corel dataset [MMMP02]. It can easily be observed that most of the images in a cluster are semantically related.

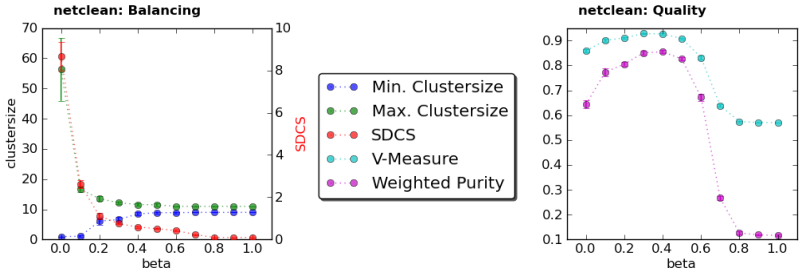

Figure 1: Results of the `bo-k-means` algorithm on the Netclean dataset. Left: Higher values of β lead to more balanced clusters. Right: A moderate balancing can improve cluster quality.

Figure 2: Example clusters of our approach on the Corel dataset. The pictures in the same cluster (displayed in the same line) are visually similar and also semantically related.

References

[AKCM90] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton. Competitive learning algorithms for vector quantization. *Neural Networks*, 3(3):277–290, 1990.

[Alt10] C. T. Althoff. *Scalable Clustering for Hierarchical Content-based Browsing of Large-scale Image Collections*. Bachelor's thesis, University of Kaiserslautern, 2010.

[BG04] A. Banerjee and J. Ghosh. Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres. *IEEE Trans. Neural Networks*, 15(3):702–719, 2004.

[BSUB08] D. Borth, C. Schulze, A. Ulges, and T. Breuel. Navidgator - Similarity Based Browsing for Image and Video Databases. In *KI 2008*, pages 22–29. Springer, 2008.

[CBD04] J. Chen, C. A. Bouman, and E. J. Delp. Vibe: A compressed video database structured for active browsing and search. In *IEEE Transactions on Multimedia*, pages 4–7, 2004.

[CWK03] Y. Chen, J. Z. Wang, and R. Krovetz. Content-based image retrieval by clustering. In *Proc. MIR 2003*, pages 193–200. ACM, 2003.

[GHP07] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology, 2007.

[HA06] J. Hofmann and M. Ali. An Extensive Approach to Content Based Image Retrieval Using Low-&High-Level Descriptors. Master's thesis, IT University of Göteborg, 2006.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review, 1999.

[MMMP02] H. Müller, S. Marchand-Maillet, and T. Pun. The Truth about Corel - Evaluation in Image Retrieval. *Image and Video Retrieval*, 2383:38–49, 2002.

[RH07] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *EMNLP'07*, 2007.