Proc. 1st ACM Workshop on Human Centered Event Understanding from Multimedia (HuEvent'14) at ACM Multimedia (MM'14), Orlando, FL, US, Nov. 2014.

Concept-based Image Clustering and Summarization of Event-related Image Collections

Christina Papagiannopoulou
CERTH-ITI
Thessaloniki, Greece
cppapagi@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece
bmezaris@iti.gr

ABSTRACT

In this work we deal with the problem of summarizing image collections that correspond to a single event each. For this, we adopt a clustering-based approach, and we perform a comparative study of different clustering algorithms and image representations. As part of this study, we propose and examine the possibility of using trained concept detectors so as to represent each image with a vector of concept detector responses, which is then used as input to the clustering algorithms. A technique which indicates which concepts are the most informative ones for clustering is also introduced, allowing us to prune the employed concept detectors. Following the clustering, a summary of the collection (thus, also of the event) can be formed by selecting one or more images per cluster, according to different possible criteria. The combination of clustering and concept-based image representation is experimentally shown to result in the formation of clusters and summaries that match well the human expectations.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Clustering; I.4.10 [**Image Processing and Computer Vision**]: Image Representation

Keywords

Image collection summarization; image clustering; model vectors; image representation; concept selection

1. INTRODUCTION

Summarization of image collections, especially collections related to an event (e.g. a social event, or a personal event such as a trip), is increasingly becoming an important topic due to the widespread availability of various digital image capture and consumption devices (e.g. smartphones, tablets, cameras) and the proliferation of media communication channels such as the Internet. Users are inundated with thousands of images, and efficient ways to organize them are necessary. The main goal of our study is the clustering and summarization of event-related image collections, in order to facilitate users to have an overview of the collections' contents and consume this summary in a wide range of related applications, from slideshow generation to digital preservation applications [9].

Image clustering has been studied in the context of several diverse applications. In order to facilitate the image retrieval task, in [14, 27] visual features were used for grouping images into meaningful clusters. More specifically, [27] uses low-level color features, such as the HSV color histogram, to represent each image, while [14] makes use of the SIFT local descriptor. Effective browsing within large image collections is the motivation of [18] and [20]; visual and textual features are combined and are used for clustering in [18], while in [20] local visual features are extracted using the SURF descriptor in order to represent the visual properties of the image, and the tags which are related with the images are used as additional features. In [29] the pixel values of the subsampled images are directly used as features for clustering. A technique which is oriented specifically towards the organization of event-related image collections is proposed in [3]: a split-n-merge algorithm is introduced, which uses the date and the time that the image was taken, the geographic coordinates of the images and the user who uploaded them on the Internet as basic features of the images.

Specifically for image collection summarization, clustering algorithms in combination with low-level visual features are often used for the selection of representative images from the collection [1, 24]. For example, at the MediaEval 2013 Retrieving Diverse Social Images task, most works use clustering methods applied to low-level visual and textual features in order to produce a summary characterized by diversity [11, 28]. In [22], textual features or user information in addition to visual features are used for composing a graph, on which the Random Walker algorithm with restart is applied to select representative images. For the summarization of social events, in [5], features such as the time, the geographical coordinates and textual information are used in order to collect and cluster the event-related images. The final selection of images for the summary is based on visual features and the number of views or likes of each image.

In this study, we compare the usage of low-level visual features for image representation, which is a technique adopted by many existing works, with the combination of clustering techniques and a concept-based approach to image representation, which builds on confidence scores provided by trained

HuEvent'14, November 07, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660505.2660507.

(a) (b)

Figure 1: Clustering examples. (a) Typical image clustering approach. The clustering algorithm is applied directly to the low-level visual features. (b) Proposed concept-based clustering approach. The clustering is performed on the responses of a pool of automatic concept detectors. These detectors are not manually selected to match the contents of the image collection, and their introduction into the clustering pipeline does not require any kind of user interaction (i.e. the entire clustering process remains fully automatic).

visual concept detectors. Although the concept detectors are far from perfect, this approach bridges to some extent the gap between the low-level representation of images and the high-level concepts that are perceived by humans when looking at them, and relies on the latter concepts for effecting a clustering and summarization that better matches human expectations. Throughout the study, we deliberately refrain from using any time or geolocation metadata that might accompany the images, so as to examine how well we can do by just considering the visual content of the images, and also to avoid being affected by any errors due to poor time synchronization across different image capture devices [2].

The rest of the paper is organized as follows. The proposed approach to image clustering and summarization is presented in Section 2. The set up of our experimental study and the extensive experiments and results that were obtained are reported in Section 3. Finally, conclusions are drawn in Section 4.

2. PROPOSED APPROACH

2.1 Overview

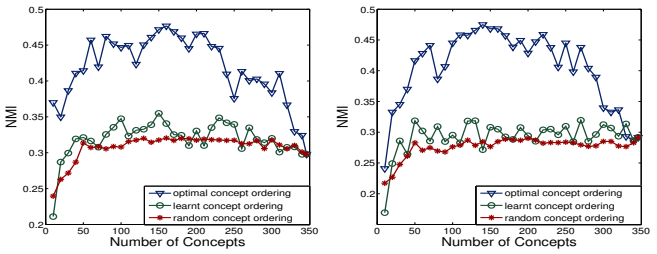
The traditional approach to clustering-based summarization has three steps. Suppose a collection I of N images $I = \{I_1, I_2, ..., I_N\}$ that we want to cluster into K clusters. Low-level visual features are extracted, which means that each image is represented as a vector of arithmetic values. These vectors are used as input to a clustering algorithm. The result is the formation of a set of clusters $G = \{G_1, G_2, ..., G_K\}$. Finally, a set of representative images $R = \{R_1, R_2, ..., R_M\}$ are selected for summarizing the image collection (often, for simplicity, one representative image per cluster, i.e. M=K). Figure 1 (a) presents the stages of typical image clustering for summarization.

In our work, we introduce a new processing stage which is inserted after the feature extraction one. Having at our disposal a pool of trained concept detectors $C = \{C_1, C_2, ..., C_J\}$, where J is the number of concepts and is typically in the order of hundreds, we apply them to the images of collection I and receive the prediction scores for each concept. Thus, each image can be represented as a J-element vector of detector confidence scores (model vectors [25]), e.g. for the i-th image the corresponding model vector has the form of $C(I_i) = [C_1(I_i), C_2(I_i), ..., C_J(I_i)]$, where $C_j(I_i)$ is the confidence score produced by the j-th concept detector C_j [25]. The confidence score is defined as the degree of confidence that a concept is depicted in the image and takes values in the range [0,1]. The clustering algorithm is then applied on these vectors and the representative images are selected as above. Figure 1 (b) shows the proposed clustering approach.

2.2 Visual Concepts for Clustering

The concept detection procedure that we used in our work is that of [15]. Three image representations (low-level descriptors) are employed for the concept detection task, based on SIFT, RGB-SIFT and OpponentSIFT. These representations are combined with two strategies for the detection of interest points, dense sampling and Harris-Laplace corner detector [8]. Additionally, two assignment methods are used in order to assign these low-level descriptors to the vocabularies created by K-means clustering, hard and soft assignment. In all cases we adopt a spatial pyramidal 3×1 scheme [12]. As a result for each combination of descriptor, point detector and assignment method, a 4000-element vector per image is formed, which is the concatenation of the 3 bag-of-words (BoW) feature vectors from the 3 horizontal bands of the image and the one which is created using the whole image (each of them has 1000 elements). Then, each such vector is used as input to the set of concept classifiers, which are linear SVMs. As for the training, we trained our framework on the TRECVID 2013 Semantic Indexing (SIN) ground-truth annotated development set [19]. The 346 concepts that we used are those defined in the TRECVID SIN task (i.e., a generic set of heterogeneous concepts, not chosen specifically to match the image collections we experimented with). Thus, each image is represented by a 346-element vector of confidence scores.

In our approach, these vectors of confidence scores are the input to the various clustering algorithms that we examine (e.g. K-means, Hierarchical). For this, the distance measure to be used during clustering has also to be defined. When working with low-level features, the squared Euclidean or the Euclidean distance, depending on the clustering algorithm, are typically used in the literature. However, for the model vectors we use an alternative distance measure which was introduced in [17], and was shown to be appropriate



(a) Image collection (k=12) (b) Image collection (k=10)

Figure 2: Curves of NMI in relation with the number of concepts.

for confidence scores comparison. According to it, if $C(I_i)$ and $C(I_k)$ are two model vectors for the images I_i and I_k respectively, the distance D of $C(I_i)$ and $C(I_k)$ is defined as:

$$D(C(I_i), C(I_k)) = \sqrt{\sum_{j=1}^{J} \frac{(C_j(I_i) - C_j(I_k))^2}{C_j(I_i) + C_j(I_k)}} \quad (1)$$

where J is the total number of concepts.

2.3 Finding a Compact Concept Bank

Inspired by the technique that was presented in [16] for finding an informative concept bank for video event detection, we adopt a similar approach for discovering the concepts which are the most useful ones for clustering. We consider that the measure that should be optimized in our case during concept selection is the same one that we use in the sequel for evaluating the clustering results, i.e. the Normalized Mutual Information (NMI) [26].

Suppose that a subset of concept detectors is denoted by x and $f(x)$ is the objective function that we want to optimize. The random variable x is characterized by the distribution $p(x;\Theta)$ where Θ is a variable that controls the importance of each concept detector to the final score $f(x)$. This random variable is modeled by the cross-entropy optimization [21]. This method has three basic steps:

- Generate n random samples according to $p(x;\Theta)$ which is modeled by a binomial distribution, i.e. each sample x is a binary vector.
- Evaluate each sample x using $f(x)$. Take the top m samples that give the best results.
- Use the m samples in order to re-estimate Θ.

This method is repeated for a number of iterations and in each iteration q the utility $\Theta_i^{(q)}$ of the concept i, taking into account the m top best samples, is calculated as:

$$\Theta_i^{(q)} = \frac{1}{m} \sum_{j=1}^{m} x_i^{(j,q)} \quad (2)$$

where $x_i^{(j,q)}$ has the value 1 if the concept i takes part to the solution and 0 otherwise. At the end of this iterative process, the higher values of Θ_i indicate more informative concepts.

2.4 Clustering-based Summarization

Following clustering, a set of images summarizing the collection needs to be selected. These are typically images belonging to different clusters, for ensuring their diversity. For the clustering algorithms that do not directly provide a representative sample for each cluster (e.g. K-means, Hierarchical), the image that is closest to the cluster's center is selected. For algorithms that indicate which member of each cluster is the closest one to all other items of the same cluster (e.g. Partitioning Around Medoids, Farthest First Traversal Algorithm) or is representative for a set of items (e.g. Affinity Propagation), we select the indicated representative images. In this way, a collection summary comprising as many images as the number of clusters is formed.

3. EXPERIMENTAL RESULTS

3.1 Datasets and Evaluation

We perform our experiments on 14 image collections, each capturing a different event. The first 8 of them are image collections of personal travel events whose size varies from 104 to 254 images, and the other 6 are image collections of social events (e.g. concerts) retrieved from the Internet and whose size varies between 159 and 325 images. For each collection two ground truth clusterings have been created manually (by one annotator for each image collection), with different numbers of clusters: one for a fixed number of clusters (k=10 for each collection), and one for a variable number of clusters (denoted k=var; the annotators were given the freedom to choose the number of clusters that would prefer for each collection). In both cases, the annotators were instructed to create the ground truth by considering the visual content of the images and how the entire collection of images of each given event could be summarized in the best way, rather than look at the time information and try to break down the collection to time-centered sub-events.

For the clustering procedure, the clustering algorithms that we studied and evaluated are: K-means [13], Hierarchical clustering using complete linkage (hier-comp) [4], Hierarchical clustering using single linkage (hier-single) [23], Partitioning Around Medoids (PAM) [10], Affinity Propagation (AP) [6] and the Farthest First Traversal Algorithm (Far. First) [7]. Three types of image representation are combined with each one of the above algorithms and are compared: low-level color features (HSV histogram), a BoW representation (created from SIFT features extracted on a dense grid and assigned to words using soft assignment), and the proposed approach, that is based on model vectors.

We evaluate the clustering results by computing the Normalized Mutual Information (NMI) [26] between each of the resulting clusterings and the manually-created ground truth. For evaluating the image collection summarization results, the Cluster Recall (CR) [30] is used. This measure allows us to assess how many clusters of the ground truth are represented in the generated summary. E.g., assuming 10 clusters in the ground truth and an automatically created summary comprising 10 images, if these 10 images belong to 7 different clusters of the ground truth, then CR(k=10) is equal to 0,7.

For the concept selection method, we set the parameters of the cross-entropy algorithm as in [16], i.e. 20 iterations, 1000 concept samples per iteration, and 200 top best samples per iteration. For the calculation of NMI during concept

Table 1: Mean and variance of the evaluation measures NMI and CR over all image collections in our dataset, separately for each possible combination of clustering algorithm - image representation.

		All event-related image collections			
image representation	clustering algorithm	NMI (k=10)	CR (k=10)	NMI (k=var)	CR (k=var)
HSV Histogram (similar to e.g. [27])	k-means	0,30 ± 0,10	0,53 ± 0,10	0,24 ± 0,11	0,56 ± 0,13
	hier-comp	0,26 ± 0,12	0,56 ± 0,14	0,19 ± 0,11	0,54 ± 0,16
	hier-single	0,18 ± 0,07	0,53 ± 0,10	0,13 ± 0,05	0,53 ± 0,15
	PAM	0,30 ± 0,11	0,56 ± 0,14	0,25 ± 0,11	0,58 ± 0,12
	AP	0,30 ± 0,10	0,59 ± 0,15	0,23 ± 0,10	0,59 ± 0,12
	Far. First	0,25 ± 0,12	0,53 ± 0,11	0,19 ± 0,09	0,63 ± 0,16
SIFT+BoW (similar to e.g. [18])	k-means	0,26 ± 0,10	0,54 ± 0,13	0,21 ± 0,09	0,52 ± 0,17
	hier-comp	0,22 ± 0,08	0,49 ± 0,11	0,17 ± 0,07	0,50 ± 0,15
	hier-single	0,16 ± 0,04	0,51 ± 0,08	0,12 ± 0,04	0,56 ± 0,10
	PAM	0,25 ± 0,09	0,51 ± 0,11	0,20 ± 0,09	0,55 ± 0,13
	AP	0,25 ± 0,09	0,50 ± 0,14	0,20 ± 0,09	0,51 ± 0,18
	Far. First	0,21 ± 0,07	0,52 ± 0,13	0,16 ± 0,07	0,51 ± 0,17
Model Vectors (all 346 concepts)	k-means	**0,36** ± 0,10	0,58 ± 0,14	0,31 ± 0,10	0,55 ± 0,17
	hier-comp	0,35 ± 0,09	0,57 ± 0,14	0,29 ± 0,10	0,56 ± 0,12
	hier-single	0,21 ± 0,10	0,54 ± 0,16	0,16 ± 0,06	0,57 ± 0,13
	PAM	0,35 ± 0,09	0,54 ± 0,13	0,28 ± 0,10	0,56 ± 0,18
	AP	0,35 ± 0,08	0,56 ± 0,11	0,28 ± 0,10	0,55 ± 0,14
	Far. First	0,30 ± 0,10	0,61 ± 0,12	0,26 ± 0,10	0,63 ± 0,17
Model Vectors (# concepts=100, selected as in section 2.3)	k-means	**0,36** ± 0,11	0,56 ± 0,07	**0,33** ± 0,13	0,58 ± 0,16
	hier-comp	0,34 ± 0,10	0,59 ± 0,11	0,30 ± 0,10	0,65 ± 0,13
	hier-single	0,21 ± 0,10	0,55 ± 0,15	0,15 ± 0,05	0,57 ± 0,12
	PAM	0,34 ± 0,10	0,57 ± 0,11	0,29 ± 0,10	0,57 ± 0,17
	AP	0,35 ± 0,11	0,57 ± 0,10	0,29 ± 0,11	0,56 ± 0,14
	Far. First	0,33 ± 0,10	0,61 ± 0,13	0,25 ± 0,10	**0,65** ± 0,17
Model Vectors (# concepts=200, selected as in section 2.3)	k-means	0,35 ± 0,10	0,54 ± 0,14	0,30 ± 0,10	0,64 ± 0,16
	hier-comp	0,35 ± 0,10	0,59 ± 0,11	0,30 ± 0,10	0,62 ± 0,13
	hier-single	0,21 ± 0,11	0,57 ± 0,12	0,14 ± 0,06	0,59 ± 0,14
	PAM	0,34 ± 0,10	0,57 ± 0,12	0,28 ± 0,09	0,57 ± 0,16
	AP	0,35 ± 0,10	0,58 ± 0,13	0,28 ± 0,09	0,58 ± 0,16
	Far. First	0,31 ± 0,10	**0,62** ± 0,11	0,25 ± 0,10	0,64 ± 0,14

selection, we use K-means as the clustering algorithm. In order to select the X most informative concepts for a given collection, we use as input to the algorithm of section 2.3 the images and ground truth clustering of all other image collections of the same type of event (i.e. personal travel, or social event) in our dataset.

3.2 Clustering Results

Table 1 presents the clustering results for all the image collections. We compute the mean value and the standard deviation of NMI over all image collections, separately for the default number of clusters (k=10), and for k=var. The last two horisontal blocks of the table present the results obtained when we apply the clustering algorithms to a subset of the total pool of concepts (100 and 200 concepts, respectively), which are selected using the learnt concept ordering for each collection (section 2.3).

As can be seen in Table 1, the model vectors in combination with the K-means algorithm give the best results (NMI). Generally, the use of model vectors gives better results in almost all cases. Specifically, using model vectors outperforms using the HSV histogram and the BoW in 134 and 158 out of 168 experiments (6 clustering algorithms × 28 ground truth clusterings of the 14 image collections), respectively. The standard deviation of NMI when using model vectors is also lower in most cases than that obtained using the HSV histograms. It should be noted that the use of Euclidean or squared Euclidean distance in model vectors would give slightly worse results than using the distance of Eq.(1). E.g. the K-means (k=10) in combination with model vectors and Eq.(1) gives mean NMI = 0,36, while using the squared Euclidean distance would result in mean NMI = 0,35. Finally, we observe that in some cases using a reduced number of concepts (i.e. 100 or 200, rather than all 346) gives better results. This can be attributed to the fact that the original pool of 346 concepts contains many concepts that are irrelevant to our dataset; pruning these concepts using the procedure of section 2.3 results in more meaningful input to the clustering algorithms.

For the concept selection technique, Fig. 2 (a) and (b) show how the NMI changes for an example image collection when we vary the number of concepts being used. For any number of concepts (shown on the horizontal axis), the exact set of concepts is selected using the procedure of section 2.3 applied on either the images and ground truth annotations for all other collections (separately for travel and social events) in our dataset (as for the Table 1 experiments; these curves are denoted "learnt concept ordering" in Fig. 2) or using the images and the ground truth clustering of the same test image collection (which gives us the maximum possible NMI that could be achieved if the concepts that are most appropriate for the given test collection were selected; these curves are denoted "optimal concept ordering"). The "random concept ordering" curve is created by randomly selecting a concept ordering (this experiment is repeated 10 times and the average results are shown in Fig. 2). The "optimal concept ordering" curves indicate that by carefully selecting a subset of the available concepts, we can have significant gains in NMI. In practice, performing this selection on different datasets does not produce such pronounced gains, but

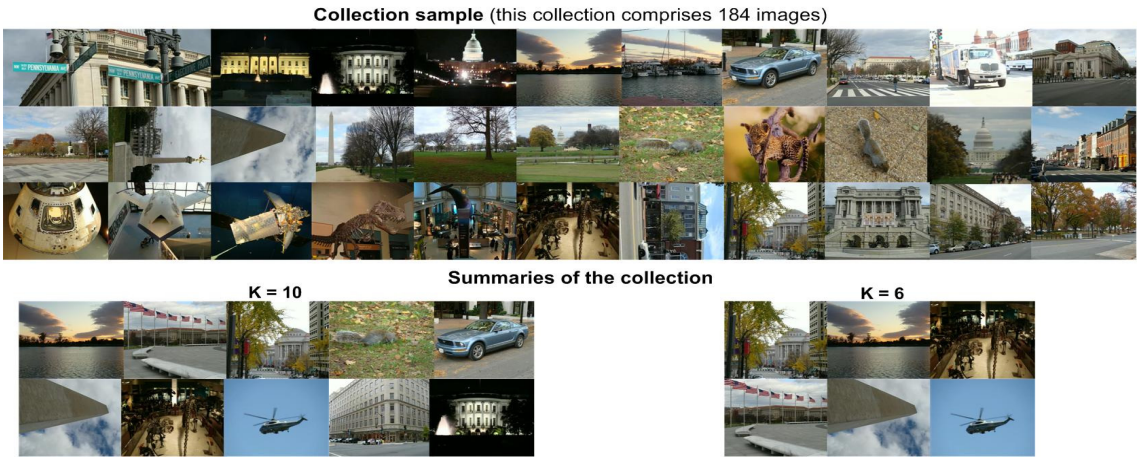


Figure 3: Summarization of an image collection of a personal travel event produced using the model vectors in combination with the farthest first algorithm.

Figure 4: Summarization of an image collection of a social event produced using the model vectors in combination with the farthest first algorithm.

does allow the reduction of the computational cost of clustering by reducing the number of concepts being used without significantly affecting the clustering effectiveness.

3.3 Summarization Results

The summarization results are also presented in Table 1. We can see that the CR evaluation measure generally takes its best value when we use the model vectors as image representation and the Farthest First Traversal Algorithm as clustering algorithm. Overall, for the image collections, the use of model vectors outperforms the use of the HSV histogram and the BoW in 113 and 128 out of 168 experiments, respectively. Figures 3 and 4 depict representative examples of the type of images in our collections and the summaries that are generated.

4. CONCLUSIONS

In this work, we examined the problem of event-related image collection summarization using a clustering approach. While most of the existing approaches in the literature use low-level visual features for image representation, we have employed concept detection confidence scores as image features. Our experimental results give evidence that the K-means algorithm combined with model vectors gives the best clustering results, and the Farthest First algorithm combined again with model vectors gives the best summarization results. Finally, we showed that selecting a subset of the available concept detectors for forming the model vectors can give us similar results to using the complete set of them, at evidently lower computational cost (since lower-dimensional feature vectors are used for clustering).

5. ACKNOWLEDGMENTS

This work was supported by the EC under contracts FP7-287911 LinkedTV and FP7-600826 ForgetIT.

6. REFERENCES

[1] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *Proc. ACM Multimedia*, pages 153–162, 2010.

[2] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of Multi-User Event Media SEM at MediaEval 2014: Task Description, Datasets, and Evaluation. In *Proc. MediaEval Workshop*, 2014.

[3] M.-S. Dao, A.-D. Duong, and F. G. D. Natale. Unsupervised social media events clustering using user-centric parallel split-n-merge algorithms. In *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing (ICASSP)*, pages 4831–4835, 2014.

[4] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[5] M. D. Fabro, A. Sobe, and L. Böszörmenyi. Summarization of real-life events based on community-contributed content. In *Proc. Fourth Int. Conf. on Advances in Multimedia (MMEDIA)*, pages 119–126, 2012.

[6] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[7] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[8] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[9] N. Kanhabua, C. Niederée, and W. Siberski. Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study. In *Proc. 10th Int. Conf. on Preservation of Digital Objects (iPRES)*, 2013.

[10] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[11] C. Kuoman, S. Tollari, and M. Detyniecki. UPMC at MediaEval 2013: Relevance by Text and Diversity by Visual Clustering. In *Proc. MediaEval Workshop*, 2013.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. CA, USA, 1967.

[14] K. Makantasis, A. Doulamis, and N. Doulamis. A non-parametric unsupervised approach for content based image retrieval and clustering. In *Proc. 4th ACM/IEEE Int. Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, pages 33–40, 2013.

[15] F. Markatopoulou, A. Moumtzidou, C. Tzelepis, and et. al. ITI-CERTH participation to TRECVID 2013. In *Proc. of TRECVID Workshop*, 2013.

[16] M. Mazloom, E. Gavves, K. V. D. Sande, and C. Snoek. Searching Informative Concept Banks for Video Event Detection. In *Proc. 3rd ACM Int. Conf. on Multimedia Retrieval*, pages 255–262, 2013.

[17] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris. On the Use of Visual Soft Semantics for Video Temporal Decomposition to Scenes. In *Proc. 4th IEEE Int. Conf. on Semantic Computing (ICSC)*, pages 141–148, 2010.

[18] P.-A. Moëllic, J.-E. Haugeard, and G. Pitel. Image Clustering based on a Shared Nearest Neighbors Approach for Tagged Collections. In *Proc. 2008 Int. Conf. on Content-based Image and Video Retrieval*, pages 269–278. ACM, 2008.

[19] P. Over, G. Awad, M. Michel, and et. al. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. TRECVID Workshop*, 2013.

[20] S. Papadopoulos, C. Zigkolis, G. Tolias, and et. al. Image Clustering through Community Detection on Hybrid Image Similarity Graphs. In *Proc. 17th IEEE Int. Conf. on Image Processing (ICIP)*, pages 2353–2356, 2010.

[21] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, 2004.

[22] S. Rudinac, A. Hanjalic, and M. Larson. Finding representative and diverse community contributed images to create visual summaries of geographic areas. In *Proc. 19th ACM Int. Conf. on Multimedia*, pages 1109–1112, 2011.

[23] R. Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[24] I. Simon, N. Snavely, and S. Seitz. Scene summarization for online image collections. In *Proc. 11th IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007.

[25] J. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, volume 2, pages II–445, 2003.

[26] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[27] C. Supriyanto, G. F. Shidik, R. A. Pramunendar, and P. N. Andono. Performance Enhancement of Image Clustering Using Singular Value Decomposition in Color Histogram Content-Based Image Retrieval. *Int. Journal of Computer and Communication Engineering*, 1(4):317–320, 2012.

[28] G. Szűcs, Z. Paróczi, and D. M. Vincz. BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information. In *Proc. MediaEval Workshop*, 2013.

[29] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 19(10):2761–2773, 2010.

[30] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development on Information Retrieval*, pages 10–17, 2003.