SEMANTIC LABELING OF MULTIMEDIA CONTENT CLUSTERS

Jelena Tešić

IBM Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
E-mail: jtesic@us.ibm.com

John R Smith

IBM Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
E-mail: jsmith@us.ibm.com

ABSTRACT

In this paper we present a novel approach for labeling clusters of multimedia content that leverages supervised classification techniques in conjunction with unsupervised clustering. Recent research has produced significant results for automatic tagging of video content such as broadcast news. For example, powerful techniques have been demonstrated in the context of the NIST TRECVID video retrieval benchmark [1]. However, the information needs of users typically span a range of semantic concepts. One of the challenges of these multimedia retrieval systems is to organize the video data in such a way that allows the user to most efficiently navigate the semantic space for the video data set. One important tool for video data organization is clustering. However, clustering results cannot be leveraged effectively when they are not labeled. We propose to build on clustering by aggregating the automatically tagged semantics. We propose and compare four techniques for labeling the clusters and evaluate the performance compared to human labeled ground-truth. We present examples of the cluster labeling results obtained on the BBC stock shots from the TRECVID-2005 video data set.

1. INTRODUCTION

The tremendous growth of multimedia content is increasing users' expectations for efficient and effective access of large multimedia repositories. However, in many cases, repositories have little or no metadata to support effective user searching, navigation and access. For example, in the domain of broadcast news, there is a category of video content called "B-rolls" or "rushes" that refers to raw or pre-production content. The volume of this type of video data is often so great, that there is little opportunity for manual indexing of the content. Similarly, with other types of video data, such as video blogs, home movies, live Web video feeds, there is often little available metadata to help to organize and index the content. This category of video data presents additional challenges for automated processing as a result of poor picture quality, tendency to be dominated by long shots with repetitive content, minimal speech, high audio noise.

The traditional approaches for video logging rely heavily on shot boundary detection or speech- and text-based indexing for organizing video data. However, to efficiently manage and discover interesting patterns, or groups of scenes in the archive, one must largely rely on the visual content. In typical scenario, a user, such as a news editor or producer, is looking for content for a story. The multimedia retrieval system needs to allow the user to efficiently navigate the semantic space of the video repository. However, the information needs of users typically span a range of semantic concepts. Modeling semantics, even the most general semantic concepts, requires investment in creating sufficient amounts of annotated video data for training the models. This is often a costly proposition. Furthermore, the space of semantics of interest to users is much larger than the space of semantic concepts that can be modeled and detected by today's systems.

The challenges presented by these large repositories requires new scalable methods that enable effective automatic organization on the scale of terabytes of video data. One important tool for video content management is clustering. Unsupervised visual clustering generally performs well for detecting redundant video content, such as when applied to repositories dominated by video rushes. However, when there is a diversity of content, the groupings are often interesting and meaningful, but still present a large space of clusters for users to navigate. Furthermore, the clustering results cannot be leveraged effectively when there is no semantic description associated with the clusters.

In this paper, we propose to build on visual clustering by aggregating automatically tagged semantics produced by concept detection techniques. The connection between visual cluster information and automatically associated semantics offers a fast and meaningful summarization of large repositories of video data. This approach enables efficient production assistance i.e. allows users to browse, search, classify, and summarize the archives without any previous knowledge of the content. We propose and compare four techniques for labeling the clusters and evaluate the performance compared to human labeled ground-truth. We analyze how well the system groups the video in topics to aid in browsing and discovery

1-4244-0367-7/06/$20.00 ©2006 IEEE

of data, and present examples of the cluster labeling results obtained on the BBC stock shots from the TRECVID-2005 video data set.

2. AUTOMATICALLY TAGGED SEMANTICS

Explicit modeling of semantics allows users to directly query the system at a higher semantic level. For example, powerful techniques have been demonstrated in the context of the NIST TRECVID video retrieval benchmark [1]. Fully-automatic approaches based on statistical modeling of low-level audio-visual features have been applied for detecting generic frequently observed semantic concepts such as indoors, outdoors, nature, man-made, faces, people, speech, music, etc. Statistical modeling requires large amounts of annotated examples for training. Since this scenarios is not feasible in the rushes archive, we adopt a new approach for automatic semantic tagging. We re-use existing semantic models, trained on the produced news and image data, to automatically associate confidence scores of rushes data with those cross-domain concept models. To enable cross-domain usability, we chose the general semantic models from LSCOM [2] lexicon, based on the consistent definitions of the concept across different image and video domains (photo albums, web, news, blogs, raw video).

3. CLUSTER LABELING

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [3]. Data clustering in low-dimensional visual descriptor space identifies and reveals most relevant trends in the visual content of the archive, and allows for summarized view of the archive. The K-means algorithm is a the simplest partitional technique that uses mean square error as a criterion. It tends to work well with isolated and compact clusters, its performance can scale well with data size [4]. However, visual clustering results cannot be leveraged effectively for the search in the semantic space if the clusters are not associated with most relevant semantic concepts. In the natural language processing, clusters are labeled with most descriptive and discriminating word using associations bigrams [5].

We propose to build on visual clustering by aggregating the automatically tagged semantics. Our approach, termed "Cluster Labeling," can be viewed as a method of inducing cluster information based on the tagged semantic of the cluster members. Information is presented as a set of concept confidences for every item. If a certain concept has a consistent scores within a cluster, and deviations of that score with respect to the other clusters is significant, assigning that particular concept to label a cluster may contain a significant amount of information. Such inductive inference of concept from a cluster, describes the content of that cluster well. We extract most descriptive visual features [6] from the representative keyframes. We associate score with each data item for each semantic concept. Then, we cluster the data items in the low-level descriptor space using K-means algorithm. To achieve the most meaningful association of a label with a cluster, we propose and evaluate 4 different statistical measures. For every method, we assign a score N_k^l that label l is relevant to cluster k.

Define X as a set of all n feature items in the archive, and let the X^l contains all the scores associated with semantic label l. Let X_k^l be defined as a set of scores for label l that belong to cluster k with n_k elements, $X_k^l = \{X_{ki}^l |\ i \in [1, n_k]\}$ and Y_k^l=$X^l - X_k^l$ as a set of scores for label l that do not belong to cluster k:

$$Y_k^l = \{Y_{ki}^l | i \in [1, \tilde{n}_k]\} \quad n = n_k + \tilde{n}_k.$$

Dominant Score A majority vote is adopted as a naive approach in this work, to enhance the dominant concept in a cluster. This approach is more resilient to estimation errors compared to the other combination strategies. Although majority rule has limited value for aggregating conflicting preferences, it offers promise for aggregating decentralized information in the cluster [7]. Let $\bar{x}_k^l$ be the mean value of scores for label l that belong to cluster k. The dominant score is:

$$N_k^l = \bar{x}_k^l \tag{1}$$

Mean Ratio The most dominant label in the cluster might not be the most descriptive one. The same dominant label might be dominant in the majority of clusters. Thus, we modify the score to reflect the relative significance of that label to all clusters, and define the mean ratio measure as,

$$N_k^l = \frac{\bar{x}_k^l}{\bar{x}^l}, \tag{2}$$

where $\bar{x}^l$ is the mean value of scores for label l over the whole dataset X^l.

T-score Student T-score [7] gives a number to whether two groups of data differ from each other in a significant way. We modify the denominator to measure the discrimination of label l score over cluster k relative to the score of label l over the rest of the data. Define vx_k^l and vx_k^l as:

$$vx_k^l = \sum_{i=1}^{n_k} \frac{(X_{ki}^l - \bar{x}_k^l)^2}{n_k(n_k - 1)} \quad vy_k^l = \sum_{i=1}^{\tilde{n}_k} \frac{(Y_{ki}^l - \bar{y}_k^l)^2}{\tilde{n}_k(\tilde{n}_k - 1)}$$

Then, the discriminative score is:

$$N_k^l = \frac{\bar{x}_k^l - \bar{y}_k^l}{\sqrt{vx_k^l + vy_k^l}} \tag{3}$$

Likelihood ratio score The likelihood ratio is a statistical measure of the goodness-of-fit between two models. We are measuring the significance of a label to a cluster with respect to its significance to other clusters. The likelihood test takes into the account probability measure associated with each score, so the scores are sigmoid normalized to fit the

[0,1] range. Likelihood for X_k^l and Y_k^l are, respectively, defined as:

$$L_{Xk}^l = \prod_{i=1}^{n_k} p(X_k^l) \quad L_{Yk}^l = \prod_{i=1}^{\tilde{n}_k} p(Y_k^l)$$

Then, the score is:

$$N_k^l = 2 * (lnL_{Xk}^l - lnL_{Yk}^l) \tag{4}$$

Given number of labels S, for every method we select the S labels l with highest score:

$$N_k^l \in MAX_j^{(S)} |N_k^j| \tag{5}$$

as the most relevant labels for the cluster k.

4. EXPERIMENTS

We analyze how well the system groups the video in topics to aid in browsing and discovery of data, and present examples of the cluster labeling results obtained on the BBC stock shots from the NIST TRECVID-2005 video data set [1]. The dataset consists of 308 video clips of vacation videos, with 19,238 extracted shots and representative keyframes.

Clusters: We extract most descriptive visual features [6] from the keyframes. We use K-means algorithm to cluster the whole dataset into 100 visually distinctive clusters in the localized color space. Localized color descriptor is extracted from a 5x5 image grid and is represented by the first 3 moments for each grid tile in the LAB color space [6].

Labels: Semantic concepts models for 13 concepts are build on the three distinct datasets: NIST TRECVID 2005 and 2003 development set [1], and IBM Team Personal Photo Annotated set. The concepts are chosen primarily based on the consistent definitions of the concept across domains, such as produced news, raw video, photo albums, video blogs, and they are: Building, Day, Desert, Greenery, Indoors, Nature, Night, Outdoors, Person, Road, Sky, Studio, and Water. Based on the concept models, semantic tags are associated with BBC shots.

Ground Truth: The ground truth was annotated by a typical user, not familiar with semantic modeling or content extraction. We asked the user to give a relevance score of how each of the 13 concepts describes the cluster. Score ranges between -1 and 1, where 1 means 'this concept describes the cluster', -1 means 'negation of this concept describes the cluster,', and 0 means 'the concept is not relevant to this cluster'. The resulting table consists of 100 × 13 entries, each entry in the range [-1,1].

Method Evaluation Experiment: We quantified the effectiveness of cluster labeling methods using normalized Mahalanobis distance measure between the sorted N_k^l scores and the ground truth. Define Err_k a cumulative error measure over number of labels S associated with a cluster k:

$$Err_k = \frac{1}{S} \sum_{i=1}^{S} |sign(GT_k(i)) - sign(N_k(i))| \tag{6}$$

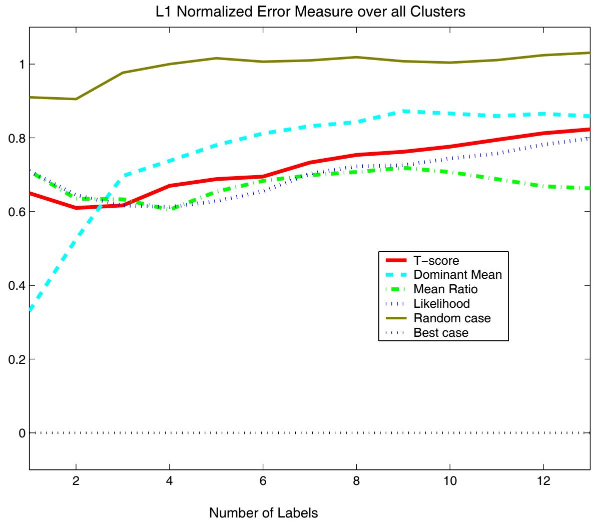
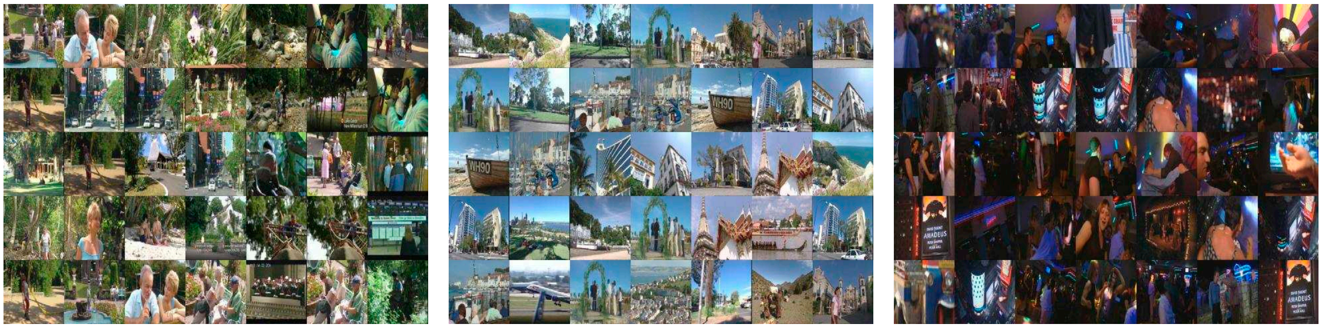


Fig. 1. Mean L_1 Error Measure for Cluster Labeling

$N_k(i)$ is the i^{th} largest method score, and $GT_k(i)$ is the ground truth for the label that produced $N_k(i)$ score for cluster k. Mean error measure is average of Err_k over all clusters, for one method. We compare Err measure for the first S labels associated with cluster k, as shown in Figure 1, $S \in [1, 13]$. When the ordering and the sign of output scores overlaps with ground truth, $Err = 0$. The worst case scenario gives $Err = 2$. Also, we randomly assign scores on the [-1,1] range and sort them according to magnitudes. As expected, the cumulative error measure is close to $Err = 1$ in this case. Note that the sign of method outputs can only be -1 or +1 while ground truth contains a lot of 0s, and this measurement system penalizes our scoring method for larger number of labels. Our methods show improvement of up to 66% than random. Dominant mean has the best performance for assigning up to top three labels to a concept, while T-score and Likelihood methods perform better in the middle range of labels: on average 40% better than random. Similar results are obtained using Euclidean measure for error.

Visual Pattern Association Experiment: We wanted to visually evaluate the relevancy and significance of the concept association with cluster. Here, we present the detailed analysis on the three out of 100 clusters. Comparison of top four labels for each of the four methods plus ground truth for all three visual clusters, together with respective N_k^l scores, is outlined in Table 1. Visual representatives of the clusters are pictured in Figure 2. Semantically distinct labels selected by T-score and likelihood emerged as good characteristics description of the distinct clusters. Dominant score method gives a cluster representative, but fails to push up the more distinctive ones. Mean score fails to capture the essence of the cluster. Overall, we find that methods that take into account both intra-concept and inter-concept dimensions select more distinct labels for a cluster.



(a) Cluster No. 10

(b) Cluster No. 33

(c) Cluster No. 58

Fig. 2. Cluster Representatives in the Local Color Space

Cluster 10	**Label 1**	**Score 1**	**Label 2**	**Score 2**	**Label 3**	**Score 3**	**Label 4**	**Score 4**
Ground Truth	NOT Night	-0.95	NOT Night	-0.95	Day	0.9	NOT Indoors	0.9
Dominant Score	Indoors	1.44	Person	1.36	Day	1.22	Outdoors	1.19
Mean Ratio	Road	1.22	Water	1.1	Nature	1.09	Desert	1.04
T-score	Person	31.28	**Nature**	24.72	NOT Desert	-21.42	Outdoors	16.4
Likelihood	**Nature**	1.2	NOT Indoors	-1.17	NOT Building	-1.09	Greenery	0.89
Cluster 33	**Label 1**	**Score 1**	**Label 2**	**Score 2**	**Label 3**	**Score 3**	**Label 4**	**Score 4**
Ground Truth	NOT Night	-0.95	NOT Studio	-0.95	Day	0.9	Indoors	-0.9
Dominant Score	NOT Studio	-1.71	NOT Night	-1.32	Day	1.29	NOT Indoors	-1.22
Mean Ratio	Indoors	1.5	Day	1.31	Sky	1.22	Outdoors	1.21
T-score	Outdoors	29.02	**NOT Indoors**	-28.66	Day	27.23	Sky	25.12
Likelihood	**NOT Indoors**	-1.53	Sky	1.07	Day	0.94	Outdoors	0.91
Cluster 33	**Label 1**	**Score 1**	**Label 2**	**Score 2**	**Label 3**	**Score 3**	**Label 4**	**Score 4**
Ground Truth	NOT Day	-1	Person	0.97	Indoors	0.85	Studio	0.8
Dominant Score	NOT Nature	-1.16	NOT Greenery	-1.05	NOT Desert	-1.03	NOT Water	-0.93
Mean Ratio	Road	1.22	Water	1.1	Nature	1.09	Desert	1.04
T-score	**Studio**	20.23	NOT Desert	-18.74	Night	15.6	NOT Sky	-12.85
Likelihood	NOT Day	-1.24	**Studio**	1.3	Night	1.02	NOT Outdoors	-0.84

Table 1. Sample Labeling Results and Associated scores for Clusters in Figure 2

5. DISCUSSION

In this paper we present a novel approach for labeling clusters in minimally annotated data archives. We propose to build on clustering by aggregating the automatically tagged semantics. We extend the usability of semantic models trained on produced news and consumer photo data, and apply them to the minimally annotated video archive. We propose and compare four techniques for labeling the clusters and evaluate the performance compared to human labeled ground-truth. We define the error measures to quantify the results, and present examples of the cluster labeling results obtained on the BBC stock shots from the TRECVID-2005 video data set. We find that methods that take into account both intra-concept and inter-concept dimensions perform better. Ground truth, labeled by one typical user might be a too subjective of a criterion. Future directions include the fusion of multiple users to obtain ground truth, and the use of more redundant set of concepts; cluster labeling using mutual label information and multi-modal fusion of cluster labeling over associated text. BBC 2005 Rushes video is copyrighted. The BBC 2005 Rushes video used in this work is provided for research purposes by the BBC through the TREC Information Retrieval Research Collection.

This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

6. REFERENCES

[1] W. Kraaij A.F. Smeaton P. Over, T. Ianeva, “Trecvid 2005 an introduction,” in *NIST TRECVID-2005 Workshop*, Gaithersburg, Maryland, 2005.

[2] S.-F. Chang, A. Haupmann, M. Naphade, and J. R. Smith, “A large scale concept ontology for multimedia understanding,” April 2005.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, vol. 31, no. 3, 1999.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., *Advances in Knowledge Discovery and Data Mining*, The MIT Press, 1996.

[5] T. Pedersen and A. Kulkarni, “Identifying similar words and contexts in natural language with senseclusters.,” in *AAAI*, 2005.

[6] A. Natsev, M. R. Naphade, and J. Tešić, “Learning the semantics of multimedia queries and concepts from a small number of examples,” in *ACM Multimedia*, Singapore, 2005.

[7] M. Berthold and D. J. Hand, Eds., *Intelligent Data Analysis: An Introduction*, Springer, 2002.