What's News, What's Not? Associating News Videos with Words

Pınar Duygulu[1] and Alexander Hauptmann[2]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey
[2]Informedia Project, Carnegie Mellon University, Pittsburgh, PA, USA
duygulu@cs.bilkent.edu.tr, alex@cs.cmu.edu

Abstract. Text retrieval from broadcast news video is unsatisfactory, because a transcript word frequently does not directly 'describe' the shot when it was spoken. Extending the retrieved region to a window around the matching keyword provides better recall, but low precision. We improve on text retrieval using the following approach: First we segment the visual stream into coherent story-like units, using a set of visual news story delimiters. After filtering out clearly irrelevant classes of shots, we are still left with an ambiguity of how words in the transcript relate to the visual content in the remaining shots of the story. Using a limited set of visual features at different semantic levels ranging from color histograms, to faces, cars, and outdoors, an association matrix captures the correlation of these visual features to specific transcript words. This matrix is then refined using an EM approach. Preliminary results show that this approach has the potential to significantly improve retrieval performance from text queries.

1 Introduction and Overview

Searching video is very difficult, but people understand how to search text documents. However, a text-based search on the news videos is frequently errorful due to several reasons: If we only look at the shots where a keyword was spoken in a broadcast news transcript, we find that the anchor/reporter might be introducing a story, with the following shots being relevant, but not the current one. A speech recognition error may cause a query word to be mis-recognized while it was initially spoken during a relevant shot, but correctly recognized as the anchor wraps up the news story, leaving the relevant shot earlier in the sequence. Expanding a window of shots around the time of a relevant transcript word may boost recall, but is likely to also add many shots that are not relevant, thereby decreasing precision. Simple word ambiguities may also result in the retrieval of irrelevant video clips (*e.g.* is Powell, Colin Powell [secretary of state], Michael Powell [FCC chairman] or the lake?).

In this paper we lay out a strategy for improving retrieval of relevant video material when only text queries are available. Our first step segments the video into visually structured story units. We classify video shots as anchors [7], commercials [5], graphics or studio settings, or 'other' and use the broadcast video editor's sequencing of these shot classes as delimiters to separate the stories.

Separate classifiers were built to detect other studio settings and shots containing logos and graphics using color features. In the absence of labeled data, these latter two classifiers were built interactively. Using color features all shots were clustered and presented to the user in a layout based on a multi-dimensional scaling method. One representative is chosen from each cluster. Cluster variance is mapped into the image size to convey confidence in a particular cluster. Clusters with low variance are manually inspected and clusters corresponding to classes of interest are selected and labeled. Story boundaries are postulated between the classified delimiters of commercial/graphics/anchor/studio shots. Commercials and graphics shots are removed. Anchor and studio/reporter images are also deleted but the text corresponding to them is still used to find relevant stories.

The final step associates the text and visual features. On the visual side, we again create color clusters of the remaining (non-delimiter) shots. In addition we also take advantage of the results from existing outdoor, building, road and car classifiers. Finally, face detection results are also incorporated, grouping shots into ones with single faces, two faces, and three or more faces. On the text side, the words in the vocabulary are pruned to remove stop words and low frequency words. Then, co-occurrences are found by counting the associations of all words and all visual tokens inside the stories. The co-occurrences are weighted by the TF-IDF formula to normalize for rare or frequent words. Then a method based on Expectation Maximization is used to obtain the final association table.

We perform standard text retrieval on the query text, but instead of expanding a global window around the location of a relevant word, the story segmentation limits the temporal region in which shots relevant to the keyword may appear. All shots within a story containing a relevant query word are then reranked based on the visual features strongly correlated to this word based on the association table. This results in clear retrieval improvement over simplistic associations between a text word and the shot where it occurred. This approach also can help to find related words for a given story segment, or for suggesting words to be used with a classifier.

Our experiments were carried out on the CNN news data set of the 2003 TREC Video Track [11]. It contained 16650 shots as defined by a common shot segmentation, with one key-frame extracted for each shot.

2 Segmenting Broadcast News Stories using Classes of Visual Delimiters

Our approach to segmentation relies on the recognition of visual delimiters inserted by the editors of the broadcasts. Specifically, we identify four types of delimiters: commercials, anchors, studio settings and graphics shots. While we had a large amount of training data for the first two delimiter types, we interactively built classifiers for the studio settings and graphics shots using a novel approach.

2.1 Identifying Commercials, Anchors, Studio and Graphics Shots

In news video broadcasts, commercials are often inserted between news stories. For efficient retrieval and browsing of news, their removal is essential, as commercials don't contain any news material. To detect commercials, we use the approach in [5], which combines analysis of the repetitive use of commercials over time with their distinctive color and audio features.

When searching for interesting broadcast news video material, detection and removal of anchor and reporter shots is also important. We use the method proposed in [7] to detect anchorpersons.

Even after the removal of commercial and anchor shots, the remaining video still contains more than the pure news story footage. Virtually all networks use a variety of easily identifiable graphics and logos to separate one story from another, or as a starting shot for a particular news category (for example characteristic logos appear before sports, weather, health or financial news) or as corporate self-identification such as the "CNN headline news" logo. While theoretically possible, it is quite tedious to manually label each of these graphics and build individual classifiers for them. However, these graphics appear frequently and usually have very distinctive color features and therefore can easily be distinguished.

In order to detect and remove these graphics, we developed a method which first clusters the non-anchor non-commercial shots using color features, and then provides an easy way to select clusters corresponding to these graphics. Similarly, shots that include studio settings with various anchors or reporters (apart from the straight anchorperson shots) can also be clustered, selected and removed to leave only shots of real news footage.

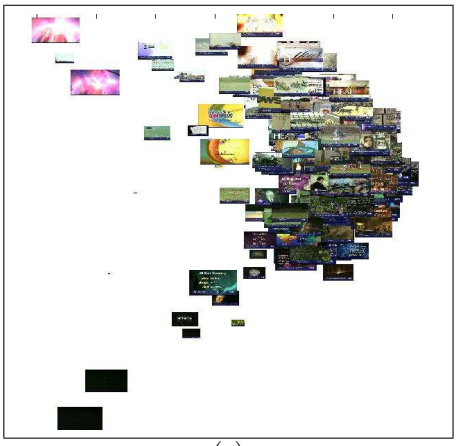
There are many ways to cluster feature sets, with differing quality and accuracy. Applying a K-means algorithm on feature vectors is a common method. However, the choice of K is by no means obvious. In this study, we use the idea of G-means [6] to determine the number of clusters adaptively. G-means clusters the data set starting from small number of clusters, C, and increases C iteratively if some of the current clusters fail the Gaussianity test (e.g., Kolmogorov-Smirov test). In our study, 230 clusters were found using color based features. The specific color features were the mean and variance of each color channel in HSV color space in a 5*5 image tessellation. Hue was quantized into 16 bins. Both saturation and value were quantized into 6 bins.

From each cluster a representative image is selected, which was simply the element closest to the mean, and these representatives are displayed using a Multi Dimensional Scaling (MDS) method. The layout is shown in Figure 1-a.

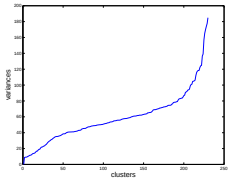
The size of the images is inversely related to the variance of the clusters. This presentation shows the confidence of the cluster. In-studio and graphics clusters tend to have less variance than other clusters, and therefore can be easily selected for further visual inspection. This process allows very quick review of the whole data set to label and remove the in-studio and graphics images. Table 1 shows the accuracy of our anchor, commercial, graphics and in-studio detection. Note that all detectors have an accuracy of 87% or higher, with commercials over

Table 1. Top: Number of shots detected and removed from each category. Remaining number of shots is 9497. **Bottom:**Number of correctly classified shots.

	anchors	commercials	graphics	in-studio
# elements	909	4347	1404	525
# correct	818 (90%)	4304 (99%)	1303 (93%)	456 (87%)



(a)



(b)

	median	mean
anchor	32.4304	32.7048
graphics	35.2399	49.8366
others	59.7816	64.7234

(c)

(d)

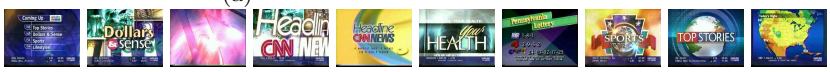


Fig. 1. (a) Representative images for the clusters. Size is inversely related to the variance of the cluster. (b) Distribution of variances for all clusters. (c) Mean and median variance values for selected graphics clusters, anchor clusters and others. (d) Example graphics clusters selected and later removed.

99%. We now remove all detected anchor, commercial, graphics and in-studio shots from the collection to leave only the shots which are related to news story footage. From the original 16650 shots, only 9497 shots were left as news story shots.

2.2 Segmenting with Visual Delimiters

To segment video news based on the visual editing structure, we devised a heuristic that uses the detected anchor, commercial, graphics and in-studio shots as delimiters. The algorithm is as follows:

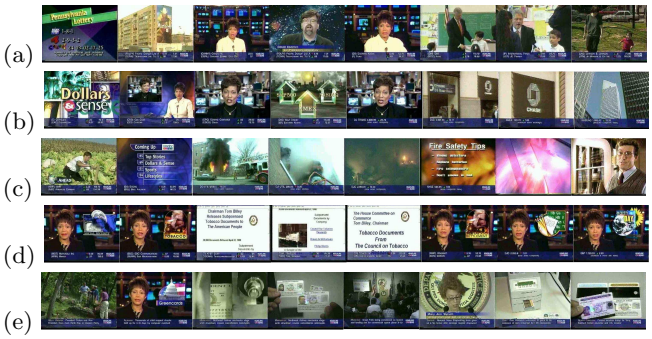
- Start a new story after a graphics or commercial shot.
- If there is a graphics or commercial in the next shot, then end the story.
- Start a new story with an anchor shot which follows a non-anchor shot.
- End a story with an anchor shot if the following is a non-anchor shot.

Figure 2 shows example segments obtained with the proposed algorithm. Most of the stories are correctly divided. Graphics create hard boundaries, while the anchor/reporter shots (and their associated text transcripts) are included

into both the preceding and following story segments. The reason is that an anchor usually finishes the previous story before starting another story. Without using textual information, the exact within-shot boundary cannot be accurately determined, nor can we tell if the anchor is only starting a story, but not finishing the previous one. Therefore, it is safer to add the anchor shots as part of both segments. This can be observed in Figure 2-d, where the iconic image in the right corner behind the anchor is same at the beginning and end of a story segment. However, in other cases, two news stories having different icons are merged into one. This problem could be solved with a more careful analysis of the icons behind the anchors or through text analysis. Other problems arise due to misclassification of the delimiter images. This may cause one story to be divided into two, or one story to begin late or end early, as in Figure 2-c and a delimiter may not be detected as in Figure 2-e . These problems again could be handled with a textual segmentation analysis.

To estimate the accuracy of our segmentation, 5 news broadcasts (1052 shots) were manually truthed for story segmentation. In the 5 broadcasts, 114 story segments were found and 69 segments were detected correctly. The errors are mostly due to incorrect splits or incorrect merges. In our case, the goal of story segmentation is to separate the news into parts for a better textual-visual association. Therefore, these incorrect segmentations actually are not very harmful or sometimes even helpful. For example, dividing a long story into two parts can be better since the further words are less related with the visual properties of the shots.

Other approaches that have been proposed for story segmentation are usually text-based. Integrated multimedia approaches have been shown to work well, however they incur great development and training costs, as a large variety of image, audio and text categories must be labeled and trained [2]. Informal analysis showed that our simple and cost-effective segmentation is sufficient for typical video retrieval queries [7].



(a) [2-3], [3-5], [5-8]
(b) [2-5], [5—]
(c) [—1], [3-5]
(d) [1-6], [6—]
(e) [1-8]

Fig. 2. Example story segmentation results. (The segment boundaries are shown as [<starting shot> - <ending shot>]. — is used to show that segment continues or starts outside of the selected shots.

3 Associating Semantics with Visual Classifiers

The video is segmented into stories and the delimiters are removed from the video as described above. The data now consists of only the shots related to the news story and the transcript text related to that story. However, the specific image/text association is still unknown. Our goal is to associate the transcript words with the correct shots within a story segment for a better retrieval.

The problem of finding the associations can be considered as the translation of visual features to words, similar to the translation of text from one language to another. In that sense, there is an analogy between learning a lexicon for machine translation and learning a correspondence model for associating words with visual features.

In [3] association of image regions with keywords was proposed for region naming and auto-annotation using data sets consisting of annotated images. The images are segmented into regions and then quantized to get a discrete representation referred as 'blob tokens'. The problem is then transformed into translating blob tokens to word tokens. A probablity table which links blob tokens with word tokens is constructed using an iterative algorithm based on Expectation Maximization [1] (For the details refer to [3]). A similar method is applied to link visual features with words in news videos [4] where the visual features (color, texture and edge features extracted from a grid) are associated with the neighbor words. However, this method have the problem of choosing the best window size for the neighborhood. In this study, we use the story segments as the basic units and associate the words and the visual features inside a story segment. Compared to [3], in our case, story segments take the place of images, and shots take the place of regions. In our study, also visual features are expanded with mid-level classifier outputs, which are called 'visual tokens'. The vocabulary is also processed to obtain 'word tokens'. The next section will give details how to obtain these tokens. Then, we describe a method to obtain the association probabilities and how they can be used for better retrieval.

3.1 Extracting Tokens

We adapt some classifiers from Informedia's TREC-VID 2003 submission [7]. Outdoor, building, car and road classifiers are used in the experiments. Outdoor and road classifiers are based on the color features explained in the previous section and on texture and edge features. Oriented energy filters are used as texture features and a Canny edge detector is used to extract edges. The classifier is based on a support vector machine with the power=2 polynomial as the kernel function. Car detection was performed with a modified version of Schneiderman's algorithm [10]. It is trained on numerous examples of side views of cars. For buildings we built a classifier by adapting the man-made structure detection method of Kumar and Hebert[8] which produces binary detection outputs for each of 22x16 grids. We extracted 4 features from the binary detection outputs, including the area and the x and y coordinates of center of mass of the bounding box that includes all the positive grids, and the ratio of the number of positive

grids to the area of the bounding box. Examples, having larger values than the thresholds are taken as building images. For faces, Schneiderman's face detector algorithm [10] is used to extract frontal faces. Here, shots are grouped into 4 different categories: no face (0), single face (1), two faces (2), three or more faces (3). Finally, color based clusters are also used after removing the clusters corresponding to in-studio and graphics shots. After removing 53 graphics and 17 in-studio clusters from 230 color clusters, 160 clusters are remained.

These classifiers are errorful. As shown in Table 2 removing the delimiters increases the accuracy of detections, but overall accuracy is very low. Our goal is to understand how visual information even if imperfect can improve retrieval results. As will be discussed later better visual information will provide better text/image association, therefore it is desirable to improve the accuracy of the classifiers, and also to create classifiers which are more specific and therefore more coherent.

On the text side, transcripts are aligned with shots by determining when each word was spoken. The vocabulary consists of only the nouns. Due to the errorful transcripts obtained from speech recognition, many incorrect words remain in the vocabulary. To remove stop words we only retained words occuring more than 10 times or less than 150 times, which cause the vocabulary to be pruned from originally 10201 words to 579 words.

Table 2. Classifier accuracies. **Before:** The original detection results on all the shots, **after:** after the removal of anchor, commercial and delimiter shots. Numbers show the number of shots detected correctly over all the detected shots. For outdoors due to the large number of images half of the data was truthed. Originally the number of detected outdoor shots was 5776 after removing anchors, delimiters and comercials.

classifier	outdoor	building	car	road
before	1419 / 4179 (34%)	126 / 924 (14%)	26 / 78 (33%)	71 / 745 (9%)
after	1000 / 2152 (46%)	101 / 456 (22%)	14 / 40 (35%)	40 / 421 (9%)

3.2 Obtaining Association probabilities

Visual tokens are associated with word tokens using a "translation table". The first step is finding the co-occurrences of visual tokens and words by counting the occurrence of words and visual tokens for each shot inside the story segments. The frequency of visual tokens are in a very large range which is also the case for words. In order to normalize the weights we apply tf-idf, which was used successfully in [9] for region-word association. After building the co-occurrence table, the final "translation table" is obtained using the Expectation-Maximization algorithm as proposed in [3]. The final translation table is a probability table which links each visual token with each word.

Figure 3 shows the top 20 words with the highest probability for some selected visual tokens. These tokens were chosen for their high word association probabilities. We observe that when a cluster is coherent the words associated

with it are closely related. Especially for sports and financial news this association is very clear. Building and road classifiers are relatively better than outdoor and car classifiers. The reason is that there are not many examples of cars, and the outdoor classifier is related to so many words due to number of outdoor shots.

The learned associations are helpful to do a better search. For a text based search, first the story segments which include the word are obtained. Then, instead of choosing the shot which is directly aligned with the query word, we choose the shot which has the highest probability of being associated with the query word. Figure 4 shows the search results for a standard time based alignment and after the words are associated with the shots using the proposed approach. For this example we choose only one shot from each story segment. Using sample queries for 'clinton' and 'fire', we find that 27 of 133 shots include Clinton using the proposed method (20% accuracy) while only 20 of 130 shots include him when the shots aligned with the text in time are retrieved (15% accuracy). For the 'fire' query, the numbers are 15/38 (40%) for the proposed approach and 11/44 (25%) for the time based approach.

stock, wall, market, street, investor, report, news, business, jones, industrials, interest, deal, thanks, cnnfn, company, susan, yesterday, morris, number, merger

series, bull, jazz, playoff, game, conference, final, karl, lead, indiana, utah, difference, combination, board, night, ball, point, pair, front, team

pilot, veteran, family, rescue, foot, effort, crew, search, security, troop, fact, affair, member, survivor, tobacco, field, department, health, communication, leader

company, market, line, worker, street, union, profit, wall, cost, news, strike, yesterday, rate, quarter, stock, check, report, level, fact, board

Fig. 3. for three color tokens and for the building token, some selected images and the first 20 words associated with the highest probability.

Fig. 4. Search results **left:** for 'clinton', **right** for 'fire'. **Top:** Using shot text, **bottom:** the proposed method. While, the time based alignment produces unrelated shots (e.g anchors for clinton), the proposed system asociates the words with the correct shots.

4 Conclusion

Association of transcripts with visual features extracted from the shots are proposed for a better text based retrieval. Story segmentation based on delimiters, namely anchor/commercial/studio/graphics shots is presented for extracting the semantic boundaries to link the shots and text. It is shown that by removing the delimiters and finding the associations it is possible to find the shots which actually correspond to the words. This method can also be used to suggest words and to improve the accuracy of classifiers. As observed in preliminary experiments, better visual tokens result in better associations. Having more specific classifiers may provide more coherent tokens. In the future we are planning to extend this work to motion information which can be associated with verbs. In this study only the speech transcript extracted was used. Text overlays can also be used for association.

4 Acknowledgements

This work was supported by the Advanced Research and Development Activity (ARDA) under contract numbers MDA908-00-C-0037 and MDA904-02-C-0451, and by the National Science Foundation (NSF) under Cooperative Agreement No. IIS-0121641. Also, we would like to thank Jia-yu Pan, Robert Chen, Rong Yan, Henry Schneiderman and Sanjiv Kumar for letting us to use their codes for detection and clustering.

References

1. P.F. Brown and S. A. Della Pietra and V. J. Della Pietra and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation", Computational Linguistics, 19:2, 1993.
2. T.-S. Chua, Y. Zhao, L. Chaisorn, C.-K. Koh, H. Yang, H. Xu, "TREC 2003 Video Retrieval and Story Segmentation task at NUS PRIS", TREC (VIDEO) Conference, 2003.
3. P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", ECCV 2002.
4. P. Duygulu and H. Wactlar "Associating video frames with text" Multimedia Information Retrieval Workshop, in conjuction with ACM-SIGIR, 2003.
5. P. Duygulu, M.-Y. Chen, A. Hauptmann, "Comparison and Combination of Two Novel Commercial Detection Methods", ICME 2004.
6. G. Hamerly and C. Elkan, "Learning the k in k-means", NIPS 2003.
7. A. Hauptmann et.al., "Informedia at TRECVID 2003:Analyzing and Searching Broadcast News Video", TREC (VIDEO) Conference, 2003.
8. S. Kumar and M. Hebert, "Man-Made Structure Detection in Natural Images using a Causal Multiscale Random Field", CVPR, 2003.
9. J.-Y. Pan, H.-J.Yang, P. Duygulu, C. Faloutsos, "Automatic Image Captioning", ICME 2004.
10. H. Schneiderman and T. Kanade, "Object detection using the statistics of parts", International Journal of Computer Vision, 2002.
11. TRECVID 2003, http://www-nlpir.nist.gov/projects/tv2003/tv2003.html