STATISTICAL MODELS FOR AUTOMATIC VIDEO ANNOTATION AND RETRIEVAL

V. Lavrenko, S. L. Feng and R. Manmatha

Multimedia Indexing and Retrieval Group
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA, 01003

ABSTRACT

We apply a continuous relevance model (CRM) to the problem of directly retrieving the visual content of videos using text queries. The model computes a joint probability model for image features and words using a training set of annotated images. The model may then be used to annotate unseen test images. The probabilistic annotations are used for retrieval using text queries. We also propose a modified model - the normalized CRM - which substantially improves performance on a subset of the TREC Video dataset.

1. INTRODUCTION

While the uses of content based video retrieval in a digital media rich world are many, it has turned out to be a challenging problem. The difficulty lies in coming up with appropriate representations for the visual content which reflects the semantics of the video. One solution has been to try to recognize objects in the videos. While some success has been achieved for objects like faces [8] much work still needs to be done to be able to recognize objects. In addition, conventional object recognition techniques require that a recognizer be trained for each object and extensive manual intervention is required to create training sets.

For NIST's TRECVID competition, it has been noted that performance based on retrieving the speech component of a news video easily exceeds any retrieval done using the visual content of the video [9, 11]. We believe that (visual) content based retrieval performance can be substantially improved using automatic annotation based techniques.

Here, we discuss two models for retrieval from images/videos based on text queries and show results on news video. Images are represented as a set of visterms i.e. each image is partitioned into a set of rectangles (usually 35 of them) and features are then computed over these rectangles. A continuous relevance model (CRM) [6, 3] represents the joint distribution of the words and visterms. The visterms are assumed to be generated by a kernel density function. The word are generated using a multinomial distribution. The difference in the two models lies in the fact that in the first model ("CRM") we assume that the annotation length may vary between images while in the second model ("normalized CRM") the annotation is of fixed length (training images may be padded with null terms to ensure this). We show retrieval experiments on annotated key frames from a subset of the TRECVID03 data and show that the second model substantially outperforms the first on retrieval using text queries. Previous annotation/retrieval experiments have mostly been performed on the Corel dataset [2, 1, 4, 6] (See [7] for an exception). It has been argued [11] that the Corel dataset is much easier than the TREC Video dataset and performance on the Corel dataset does not imply that such techniques will work on real world datasets like those in TREC Video. Our results here are comparable to those achieved for the Corel dataset and show that annotation models are applicable to such tasks (see also [7]).

The rest of the paper is laid out as follows. The next section describes the models used here and their connections to related work. This is followed by a section on experimental results on a subset of the TREC Video collection.

2. MATHEMATICAL FRAMEWORK

2.1. Overview of the Model

The *Continuous Relevance Model* CRM[6] is a statistical model for automatically assigning keywords to unlabeled images. The model relies on a training set of annotated images and operates as follows. First, we partition each training image into regions (either using an unsupervised segmentation algorithm or by partitioning the image into rectangular regions). Then, we compute a real-valued feature vector for each region. The features may include shape,

This work was supported in part by the Center for Intelligent Information Retrieval, by the National Science Foundation under grant NSF IIS-9909073 and by SPAWARSYSCEN-SD under grant N66001-02-1-8903.

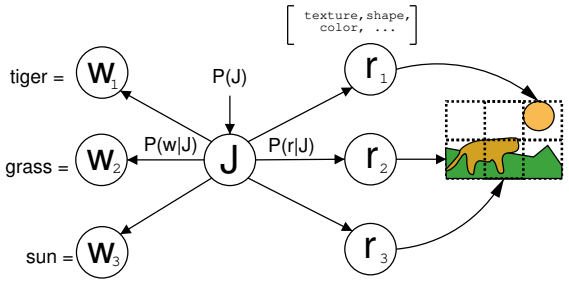


Fig. 1. CRM as a process for generating annotated images. First, pick a training image J. Then, sample the annotation words $w_1 \ldots w_m$ from the multinomial distribution $P(w|J)$. Finally, sample image regions $r_1 \ldots r_n$ from the density function $P(r|J)$.

color and texture of a region. As a result, each training image is represented as a set of feature vectors $\mathbf{r} = \{r_1 \ldots r_n\}$ along with a set of annotation words $\mathbf{w} = \{w_1 \ldots w_m\}$. As a final step, we construct a joint probability distribution $P(\mathbf{w}, \mathbf{r})$ over the annotation words $\mathbf{w}$ and image features $\mathbf{r}$. This joint distribution allows us to find most likely annotations for new unlabeled images by searching for words $\mathbf{w}$ that maximize the conditional probability $P(\mathbf{w}|\mathbf{r}) = P(\mathbf{w}, \mathbf{r})/P(\mathbf{r})$.

Unlike other annotation models the *translation* model[2], and the *correspondence LDA* model[1], CRM[6] uses a doubly non-parametric technique in computing $P(\mathbf{w}, \mathbf{r})$. The probability is computed as joint expectation over the space of distributions $P(\cdot|J)$ defined by annotated images J the training set $\mathcal{T}$:

$$P(\mathbf{w}, \mathbf{r}) = \sum_{J \in \mathcal{T}} P(J) P(\mathbf{w}, \mathbf{r}|J) \qquad (1)$$

By relying directly on individual images in the training set, the CRM allows the data to speak for itself and avoids making a-priori assumptions about the structure of the space (e.g. the latent aspect assumption in [1]). In addition, CRM makes no assumptions about the alignment of annotation words to image regions (as was done in [2]). Instead, CRM assumes that annotation words $w_1 \ldots w_m$ and image regions $r_1 \ldots r_n$ are all conditionally independent given the training image J:

$$P(\mathbf{w}, \mathbf{r}|J) = \prod_{w \in \mathbf{w}} P(w|J) \prod_{r \in \mathbf{r}} P(r|J) \qquad (2)$$

The annotation component $P(w|J)$ of the joint probability is modeled using a multinomial distribution. The feature component $P(f|J)$ uses a non-parametric density estimate with Gaussian kernels (details are provided in 2.4). Figure 1 shows a graphical representation of the CRM and the independence assumptions inherent in the model.

2.2. Adapting CRM to the Video Domain

In the course of our experiments on the video dataset[10] we found two deficiencies in the original formulation[6] of the CRM. First, the reliance on automatic segmentation proved to be problematic both in terms of computational expense and in terms of accuracy of the model. We, therefore, use a rectangular partition (see also [3]). As we show in section 3, this results in a substantial improvement in performance. We hypothesize the improvement comes from the fact that rectangular regions result in more redundancy, and CRM is able to produce more stable associations between regions and annotation words. Second, the estimation technique for the multinomial component $P(w|J)$ turned out to be a poor choice for video annotations which are hierarchical and widely varied in length. In the remainder of this section we take a detailed look at the second problem and discuss a successful approach for dealing with it.

2.2.1. Multinomial Estimation

The original formulation of CRM[6] used a simple Bayesian estimate for the annotation probability $P(w|J)$. The estimate was based on the relative frequency of the word w in the annotation of image J:

$$P(w|J) = \lambda \frac{N_{w,J}}{N_J} + (1 - \lambda) \frac{N_w}{N} \qquad (3)$$

Here $N_{w,J}$ is the number of times w occurred in the annotation of J, N_J is the length of annotation, N_w is the total number of times w occurred in the training set, and N is the aggregate length of all training annotations. λ denotes a parameter that controls the degree of smoothing.

The estimate provided by equation (3) reflects the *prominence* of the word w in the annotation. For example, if some image J_1 is annotated with a single word "face", then (barring the effects of smoothing) we get $P(\text{face}|J_1) = 1$. If some other image J_2 is annotated with ten different words, one of which is "face", we get $P(\text{face}|J_2) = \frac{1}{10}$. Arguably, both images contain "face" in their annotations, so the probabilities should not differ by an order of magnitude. To make the probabilities more comparable we expanded all annotations to a fixed length $N^* = \max_J\{N_J\}$. This was accomplished by adding $(N^* - N_J)$ instances of a special *"null"* word to the annotation of image J. We refer to this variation of the model as *normalized CRM* and demonstrate that it achieves substantially better retrieval performance than the original CRM.

2.3. Using the model for Video Retrieval

We use a retrieval model based on the *language modeling* approach to information retrieval, which was pioneered

by Ponte and Croft[5] and later adopted by numerous researchers. The idea behind the language modeling approach, is to estimate a language model M_D for every document D in the collection, and then rank the documents by the probability of observing the query $\mathbf{w}_Q$ from the model of each document. Specifically, the documents D are ranked by:

$$P(\mathbf{w}_Q|M_D) = \prod_{w \in \mathbf{w}_Q} P(w|M_D) \quad (4)$$

In our case "documents" correspond to key frames I in the video. The language model M_I can be computed as follows. Let $\mathbf{r}_I$ denote the set of feature vectors for image I. Equation (1) gives us a way to compute the joint probability $P(w, \mathbf{r}_I)$ for every word w in the vocabulary. Once that is done, we can marginalize the joint distribution to get the desired language model: $P(w|M_I) = P(w, \mathbf{r}_I)/P(\mathbf{r}_I)$. This language model can be used in equation (4) to rank video key frames in response to the textual query $\mathbf{w}_Q$.

2.4. Feature Modeling Details

In the earlier discussion we omitted the details of the feature component of our model. $P(r|J)$ is a density function responsible for modeling the d-dimensional feature vectors $r_1 \ldots r_n$, which are computed from the rectangular regions of each image. We use a non-parametric kernel-based density estimate for the distribution $P(r|J)$. Let $\mathbf{r}_J = \{r_1 \ldots r_n\}$ be the set of regions of image J. We estimate the probability density for a new vector r as:

$$P(r|J) = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp\left\{(r - r_i))^\top \Sigma^{-1} (r - r_i))\right\}}{\sqrt{2^d \pi^d |\Sigma|}} \quad (5)$$

Equation (5) arises out of placing a Gaussian kernel over the feature vector r_i of every region of image J. Each kernel is parametrized by the feature covariance matrix Σ. As a matter of convenience we assumed $\Sigma = \beta \cdot I$, where I is the identity matrix. β plays the role of kernel *bandwidth*: it determines the smoothness of $P(r|J)$ around the support points r_i. The value of β is selected empirically on a held-out portion of the training set $\mathcal{T}$.

3. EXPERIMENTAL RESULTS

In this section, we provide the experimental results of the retrieval task over a key frame dataset, which is a subset of NIST's TREC Video dataset. [1]. The data set consists of 12 mpeg files, each of which is a 30-minutes video section of CNN or ABC news and advertisements. 5200 key frames were extracted and provided by NIST for this dataset. The participants in TREC Video annotated a portion of the videos. The word vocabulary for human annotation is represented as a hierarchical tree with each annotation word as a node. For example, a key frame can be assigned a set of words like "face, female_face, female_news_person". This means that the annotation length for key frames can vary widely. There are 137 keywords in the whole dataset after we ignore all the audio annotations.We randomly divide the dataset into a training set (1735 key frames), a validation set (1735 key frames) and a test set (1730 key frames). The validation set is used to find system parameters, and then merged into the training set after we find the parameters.

Query length	1 word	2 words	3 words
Number of queries	107	431	402
Relevant images	6649	12553	11023
Precision at 5 retrieved key frames			
CRM	0.36	0.33	0.42
normalized CRM	0.49	0.47	0.58
Mean Average Precision			
CRM	0.26	0.19	0.25
normalized CRM	0.30	0.26	0.32

Every key frame in this set is partitioned into rectangular grids, and a feature vector is then calculated for every grid region. The number of rectangles is empirically selected (using the training and validation sets), which is 35 for each sample. The feature set consists of 30 features: 18 color features (including region color average, standard deviation and skewness) and 12 texture features (Gabor energy computed over 3 scales and 4 orientations). Previous results have shown that CRM based on rectangles outperforms that based on segmentation substantially, as on the Corel[3] dataset we got 0.235 average precision on CRM using segmentation, and 0.303 on CRM with rectangles for 179 one-word queries.

For the video retrieval task, each video shot is usually represented by a key frame. So we could retrieve video shots through the retrieval on their corresponding key frame set. Given a text query and a collection of un-annotated key frames, then our goal is to return all the relevant key frames, ranked according to the probabilities obtained using our retrieval model. In our retrieval experiments, we use three sets of queries [2] constructed from all 1-, 2-, 3- combinations of words which occur at least 10 times in the testing set. For each set of queries, we do comparative experiments using both CRM and normalized CRM. An image is considered relevant to a given query if its manual annotation contains all the query words. Evaluation metrics are precision at 5 retrieved key frames and non-interpolated average precision, averaged over the entire query set.

[1] We used a subset because of the computational requirements of the algorithm. We hope that in the future, faster processors as well as more effi cient implementations will allow us to use larger datasets

[2] Given that we used only a subset of the TREC Video it did not make sense to use TREC Video queries

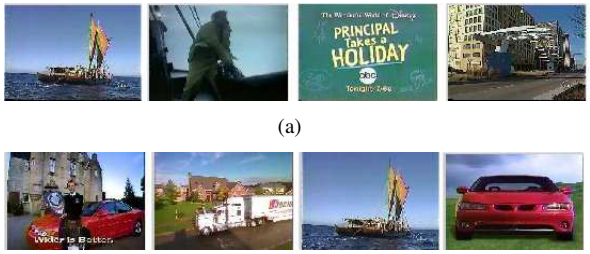
(a)

(b)

Fig. 2. First 4 ranked results for the query "Basketball".

Table 3 compares the performance of CRM and normalized CRM for three sets of queries. We observe that the latter method outperforms the former substantially, by 15%, 37% and 33% on the 1-, 2- and 3-words query sets respectively. For all three query sets the differences in precision are statistically significant at the 1% level according to the sign test. The precision at 5 retrieved key frames also indicates that using the latter method, there are half images relevant to the query in the top 5. We observe that normalized CRM consistently outperforms CRM over all query sets.

Figures 2 and 3 show the top 4 images in rank order using both CRM and normalized CRM corresponding to the text queries " basketball" and "outdoors sky transportation" respectively. We see that normalized CRM does much better than CRM.



(a)

(b)

Fig. 3. First 4 ranked results for the query "Outdoors, Sky, Transportation".

4. CONCLUSION

We have shown that statistical models can be used to retrieve video by content. Future work will include improvements to the model's speed so that it can be tested on larger datasets.

5. REFERENCES

[1] D. Blei, and M. I. Jordan. (2003) Modeling annotated data. In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 127–134, 2003

[2] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conf. on Computer Vision*, pages 97-112, 2002.

[3] S. L. Feng, V. Lavrenko and R. Manmatha. Multiple Bernoulli Relevance Models for Image and Video Annotation. Submitted to CVPR'04.

[4] J. Jeon, V. Lavrenko and R. Manmatha. (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 119–126, 2003

[5] Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. Proceedings of the 21st Intl. ACM SIGIR Conf., pages 275–281.

[6] V. Lavrenko, R. Manmatha and J. Jeon. A Model for Learning the Semantics of Pictures, To appear in Proceedings of NIPS'03.

[7] P. Duygulu and H. Wactlar Associating video frames with text In Proceedings of the SIGIR Multimedia Information Retrieval Workshop 2003, Aug, 2003. Also at http://km.doc.ic.ac.uk/mmir2003/

[8] H. Schneiderman, T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. Proc. IEEE CVPR 2000: 1746-1759

[9] A. F. Smeaton and P. Over. The TREC-2002 video track report. In *E. M. Voorhees and D. K. Harman, editors, The Eleventh Text REtrieval Conference (TREC-2002)*

[10] http://www-nlpir.nist.gov/projects/trecvid

[11] T. Westerveld and A.P. de Vries. Experimental Evaluation of a Generative Probabilistic Image Retrieval Model on 'Easy' Data. In Proceedings of the SIGIR Multimedia Information Retrieval Workshop 2003, Aug, 2003. Also at http://km.doc.ic.ac.uk/mmir2003/