SEMANTIC ANNOTATION OF MULTIMEDIA USING MAXIMUM ENTROPY MODELS

Janne Argillander, Giridharan Iyengar, Harriet Nock

IBM TJ Watson Research Center
Yorktown Heights, NY 10598
Email: janne@fi.ibm.com,{giyengar,hnock}@us.ibm.com

ABSTRACT

In this paper we propose a Maximum Entropy based approach for automatic annotation of multimedia content. In our approach, we explicitly model the spatial-location of the low-level features by means of specially designed predicates. In addition, the interaction between the low-level features is modeled using joint observation predicates. We evaluate the performance of semantic concept classifiers built using this approach on the TRECVID2003 corpus. Experiments indicate that our model performance is on par with the best results reported to-date on this dataset; Despite using only unimodal features and a single approach towards model-building. This compares favorably with the state-of-the-art systems which use multimodal features and classifier fusion to achieve similar results on this corpus.

1. INTRODUCTION

Growing amounts of multimedia content, especially video have reached a critical point where methods for indexing, searching, and efficient retrieval are expressly needed to manage the information load. The amount of multimedia content that is already present in most consumer hard-drives makes manual annotation (and consequently, indexing using high-level keywords) impossible. There has been some effort in using query-by-example (QBE) to seek into multimedia content (e.g [1, 2] amongst many others). While QBE is a powerful paradigm, its reliance on low-level perceptual properties is counter to the *semantic* nature of most user queries[1, 3]. Query-by-keyword (QBK) systems, where the user queries the content using semantic descriptors, are getting more and more attention. These systems typically require at least two processing stages: A primary *training* phase where the system is taught to identify specific concepts from a pre-defined vocabulary; A secondary *annotation* phase where the system (semi-)automatically annotates previously unseen content with these newly learned concepts.

In this paper, we focus on our recent work in using Maximum Entropy (MaxEnt) modeling techniques for automatic annotation of multimedia content. In our particular approach, the problem is formulated similar to a multiple instance learning problem. By this we mean that the annotations are specified at the level of the entire image or video shot. That is, given an annotation such as *face*, we know that the particular training shot contains a face but not its precise location within the shot. This is aimed at reducing the *acquisition effort* involved in training these semantic concept models[1].

[1]See our previous discussion on the three distinct dimensions along which we believe concept modeling systems need to be measured[4].

The rest of the paper is organized as follows: In section 2 we provide a quick summary of related work in this area. The details of our MaxEnt modeling approach follows in section 3. The dataset and experiments are detailed in section 4 followed by conclusions.

2. RELATED WORK

There is extensive literature in object detection (especially human faces) where the extent of an object is well-marked in an image. There is relatively limited literature in automatic image annotation where the physical extent of objects are not specified. In one set of approaches, techniques from statistical machine translation were applied to the problem of image annotation. In these approaches it is assumed that the annotation and the associated image are translations of each other and with a suitable of *tokenization* of the image features, standard machine translation models have been applied with some success[5]. Motivated from a cross-lingual information retrieval perspective, Lavrenko et al.[6] approach image annotation as an example-based learning problem where perceptual similarity in the image space is assumed to generate similar annotation words. Both these approaches have been demonstrated on relatively small datasets (5000 images from COREL dataset) and they remain to be evaluated in larger contexts such as what is attempted in this paper (e.g. 80000 shots from TRECVID2003 corpus[7]). Motivated by the under-constrained nature of the annotation problem together with the non-independent nature of low-level image features, we approach this in a Maximum Entropy setting which has had remarkable success in many Natural Language Processing tasks such as sentence-boundary detection and parts-of-speech tagging[8, 9]. A similar approach using MaxEnt for image annotation was proposed in[10]. The novelties of our approach are two-fold: We model the spatial- and joint-dependence between low-level features using specially designed predicates. We believe such information is important for objects that have a well-defined spatial composition (e.g. faces). In addition, we evaluate our approach on a much larger corpus (TRECVID2003). Furthermore, we present a comparison of our approach with previously published results on the TRECVID2003 concept detection task[7, 11].

3. MAXIMUM ENTROPY APPROACH FOR MULTIMEDIA ANNOTATION

In MaxEnt modeling, we assume that a random process produces an output (label) y given a context x. In multimedia annotation, y, which is a member of a finite set (vocabulary) Y, can be seen as a label for a specific shot. And x, a member of a finite set X,

as extracted information (features) from the current frame. Training data is presented in pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. The task is to learn possible correlations between x and y, and to build statistical models that can be used to annotate previously unseen shots automatically. The empirical probability distribution function (pdf) based on training data is as follows

$$\tilde{p}(x, y) = \frac{1}{n}\text{freq}(x, y) \tag{1}$$

Where *freq* is the count of a specific pair (x, y) in the training data. In real world applications, the training set size is finite. Therefore, the empirical distribution is a poor estimate of the joint pdf. Based on this partial information, MaxEnt modeling can be used to estimate the pdf that generated the empirical distribution $\tilde{p}(x, y)$ in an unbiased way[12]. At the core of the modeling process are feature functions. In this paper we prefer the term *predicates* over *feature functions* to avoid confusion with extracted low-level image features. These predicates are used to specify constraints on the model. In MaxEnt, the process of defining predicates is central to modeling: The goodness of the models is dependent on the ability of these predicates to capture relevant information. We now detail the different predicates used to capture a variety of spatial and co-occurrence properties of the low-level image features. We note again that this is a fundamental difference between our approach for multimedia annotation over previous work[10].

In our experiments, we extract 3 types of low-level image features from each video shot: Lab space color moments (mean, variance, skewness and kurtosis for each channel), Edge orientation histogram (Edge strength and orientation values at each pixel, each quantized to 8 bins) and summary statistics of grey-level co-occurrence matrices (entropy, energy and contrast values). Together, these form our 3 different low-level descriptors which we will term *Color*, *Edge* and *Texture* in further discussions. Furthermore, we partition each shot key-frame (comprising 350×240 pixels) into 35 regions (50×48 pixels each) and extract the feature descriptors for each of these 35 regions.

3.1. Unigram predicates

Unigram predicates are defined to capture the co-occurrence statistics between a specific tokenized descriptor and manual annotation of the training data. All unigram predicates used in this paper have following form:

$$f_{cd^i,a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i \in x^i, i = 0, 1, 2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

A predicate of this type is active only if tokenized descriptor cd is in current frame x and the corresponding manual annotation is a. The total number of unique unigram predicates in our model is (descriptor count x cluster size) 3x25=75.

3.2. Place Dependent Unigram Predicates

Place dependent unigram predicates are designed to capture location specific statistics. For instance, these predicates help the model learn that regions corresponding to *sky* are usually in upper parts of a key-frame.

$$f_{cd^i,a}(x^i_r, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i = x^i_r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Where region r has values $0..34$ and descriptor i takes values $0, 1, 2$. The place dependent predicate is active only if the tokenized descriptor cd is in region r of the current frame which has the annotation a. The total number of such predicates in our model is (descriptor count x region count x cluster size) 3x35x25=2625.

3.3. Bigram Predicates

In our work we have experimented with two types of bigram predicates:horizontal and vertical. These predicates model the relationship between neighboring regions. Below is an example horizontal bigram predicate which is active only if tokenized descriptor cd_r and its horizontal neighbor cd_{r+1} is adjacent in current frame x with annotation a.

$$f_{cd^i_r+cd^i_{r+1},a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i_r + cd^i_{r+1} \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where the region r take values so that the adjacent region on right cd^i_{r+1} is in the row. The following equation illustrates a vertical bigram predicate which is active only if tokenized descriptor cd_r and its vertical neighbor cd_{r+7} are also adjacent in current frame x.

$$f_{cd^i_r+cd^i_{r+7},a}(x^i, y) = \begin{cases} 1 & \text{if } y = a \text{ and } cd^i_r + cd^i_{r+7} \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Where the region r take values so that the adjacent region below is in same column $r = 0..27$.

Both types of bigrams are constructed by combining the tokenized features in the product space of the unigram predicates. This choice imposes the possibility of obtaining bigram values that are not supported in the training data, resulting primarily from the sparseness of the product space. To counter this, we employ an approach inspired from class-based language models in speech processing. When two unigrams are composed into a bigram, we treat them differently. We start with few clusters for the composed unigrams and slowly increase the number of clusters such that the number of unique bigram predicates observed (in the training data) at each step matches the total possible bigram product space values. We stop at the largest cluster size for which this condition is met in the training data.

3.4. Joint Observation Predicates

The predicates discussed so far model individual low-level feature descriptors (i.e. Color, Edge, Texture). We now illustrate predicates that model the interactions between the various low-level feature descriptors.

$$f_{cd,a}(x, y) = \begin{cases} 1 & \text{if } \forall i \; y = a \text{ and } cd^i \in x^i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This joint observation predicate is active only if all low-level descriptors are present in a given region. In our experiments we work with 144 such joint observation predicates, chosen using validation data. Figure 1 illustrates the various types of predicates used in our model.

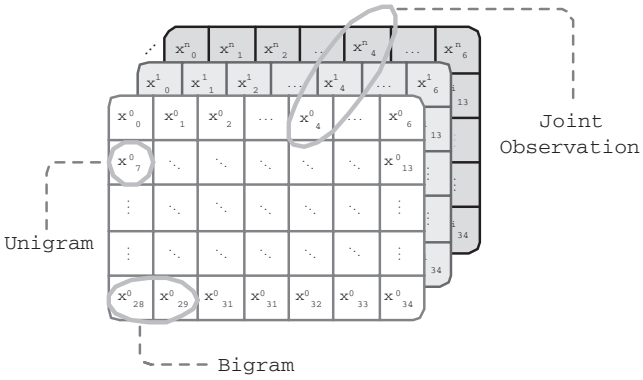

Fig. 1. *The figure illustrates the 35-region grid partition of shot key-frames. Shown in the figure are the place-dependent unigrams, horizontal bigrams and joint-observation predicates.*

3.5. Model Preparation

In this work, we built a distinct binary classifier for each semantic concept that we used to evaluate the models. Keeping with the Multiple Instance learning approach, each labeled example containing the target concept annotation[2] and is considered a positive example for those classifiers. All training instances that do not contain the target annotation are marked as negative examples. These empirical expectations of the various predicates provide constraints for the MaxEnt modeling. We use the open source MaxEnt modeling toolkit[13] in our experiments and in particular use the Generalized Iterative Scaling (GIS) algorithm[14] with smoothing to solve for the conditional probability density function.

3.6. Automatic Annotation of Unseen Multimedia Content

Semantic concept annotation of unseen multimedia content proceeds in the following manner. First, the low-level feature descriptors are extracted from the data, using the same 35-grid partitioning of the shot key-frames. For each concept to be predicted, the set of all active predicates relevant to the concept are extracted from the feature descriptors. We now have enough information to estimate the conditional probability of the particular annotation for the shot key-frame.

4. DATASET AND EXPERIMENTS

We now detail the dataset used to evaluate the MaxEnt models for semantic concept annotation of multimedia data and present our experimental results.

4.1. Dataset and model preparation

We use the TRECVID2003 corpus comprising 120 hours of broadcast news videos for our experiments. This corpus is further divided into approximately evenly sized test and development partitions. We compare the performance of our system with the NIST-evaluated relevance judgments reported on the test partition. For the development partition NIST has provided ground truth annotations at the video-shot level. In addition, NIST has provided reference key-frames for each shot for the entire corpus. For each of these reference key-frames we extract the specified low-level feature descriptors on a 35-grid layout as indicated earlier in the paper. We associate the shot-level ground truth annotations to each of the reference key-frames in development partition. We note that these annotations are provided at the shot-level and do not specify spatial or temporal boundaries of objects within a shot (i.e. we know that a face appeared in the shot but we do not know *when* and *where* within this shot). We selected 12 of the 17 benchmarked concepts from TRECVID2003; We removed audio concepts (*Female Speech*), abstract concepts (*Physical Violence*), specific person concepts (*Madeleine Albright*), camera operations (*Zoom-in*) and multimodal concepts (*News Subject Monologue*). In the case of audio and multimodal concepts, these were removed because our low-level features do not capture relevant information. Camera operations do not belong in the same category of concepts as the rest of the concepts. Both *Physical Violence* and *Madeleine Albright* had very few training examples. We note that the performance of the benchmark systems on these concepts were quite low as well. For each of the selected concepts, we build a MaxEnt classifier as previously stated. These trained classifiers are then used to annotate the test corpus.

[2]We note here that each training example has multiple concept labels in its ground truth annotation. E.g. A shot may be annotated as a face, outdoors, sky etc.

4.2. Results

NIST provides pooled relevance judgments and since our system was not part of this pooling, it would be unfairly biased to compare our system with the pooled judgments. To make comparisons valid, we choose to evaluate the different systems using precision at the top 100 retrieved shots as opposed to the average precision metric that is used by NIST. Furthermore, we restrict our comparison to two of the top 10 semantic concept detection systems at TRECVID2003: the best performing (multimodal) system and the best unimodal system[11]. All results are detailed in Table 1. The table also details the 12 concepts classifiers that we built. The first column BOU (Best Of Unimodal) is formed from the set of models by selecting the best performing unimodal classifier for the semantic concept under consideration. For instance, the best unimodal *weather* classifier could have been based on the speech recognizer output and not on visual features. The second column BOBO (Best Of Best Of) is the primary run submitted by IBM at TRECVID2003[11]; And it represents the best multimodal model including information fusion across modalities and classifier fusion across different classifiers. For further details on this system, please refer to our TRECVID2003 description[11]. The third column shows results using MaxEnt modeling approach detailed in this paper. A sample result showing the top 12 retrieved matches for *News Subject Face* is illustrated in Figure 2.

From the results we see that MaxEnt out-performed BOU in 7 concepts and BOBO in 5 concepts. We note here that in the case of BOU and BOBO, the systems had access to a commercial detector and this was used to selectively improve the concept detectors[11]. In addition, in the case of BOU, the choice of modality (i.e. audio, text or visual information) and granularity of feature extraction (global versus regional) varied across the different concepts, based on performance on a validation set. In the case of BOBO, the variation spanned not just on input modalities and granularity but also on modality fusion and classifier fusion techniques employed in the final model. On the other hand, the MaxEnt models

Concept	BOU	BOBO	MaxEnt
Outdoors	0.81	0.85	0.98
News Subject Face	0.80	0.73	0.94
People	0.90	0.99	0.92
Building	0.53	0.56	0.55
Road	0.46	0.52	0.67
Vegetation	0.96	0.93	0.91
Animal	0.10	0.10	0.11
Car Truck or Bus	0.68	0.56	0.63
Aircraft	0.38	0.63	0.32
Non Studio Setting	0.97	0.97	0.96
Sports Event	0.81	0.98	0.94
Weather	0.81	0.98	0.68
Mean Precision	**0.68**	**0.73**	**0.72**

Table 1. Results of the MaxEnt models compared against the TRECVID2003 benchmark system results. The numbers are Precision at 100 retrieved shots.

Fig. 2. *The top 12 results using the MaxEnt models for the News Subject Face concept. Note that the statue is an incorrect classification for this concept.*

rely only on the visual features and operate on a fixed feature granularity across all evaluated concepts. We further note that a signed t-test between BOU and BOBO indicates significance only at the 90% confidence level and the differences between the MaxEnt and BOBO approaches are not statistically significant.

5. CONCLUSIONS

In this paper we detailed a Maximum Entropy approach for automatic semantic annotation of multimedia data. This approach was evaluated on the TRECVID2003 corpus and benchmarked against the top ranked systems. The results indicate that this approach is promising and performs as well as the state-of-the-art multimodal systems for automatic semantic annotation despite using a single feature modality. This is a very encouraging result. Further study is needed to evaluate the effect of feature granularity selection (e.g. we posit that concepts such as *weather* and *outdoors* will benefit from global features) and more importantly, inclusion of other modalities (such as audio and speech. E.g. *weather* has a distinct vocabulary) on the performance of the MaxEnt models.

In addition, we have built multiple binary classifiers in these experiments. A natural question to address is the performance difference between a single multi-way MaxEnt classifier for all concepts versus multiple binary classifiers. In addition, we note that the ground truth annotations can be quite variable in quality; Frequently, the common objects are not marked. For instance, concepts such as *Outdoors* tend to be missing in many annotations despite being present in the shot. This needs to be explicitly accounted for by allowing the model to handle *unlabeled* objects and regions in a video shot. We intend to address these issues in a future paper.

6. REFERENCES

[1] John R. Smith and Shih-Fu Chang, "VisualSEEk: A fully automated content-based image query system," in *Proceedings of the ACM Multimedia Conference*. ACM, 1996, pp. 87–98.

[2] M. Flickner, H. Sawhney, and et al, "Query by image and video content: The qbic system," *IEEE Computer*, vol. 28(9), pp. 23–32, 1995.

[3] Martin Szummer and Rosalind W. Picard, "Indoor-outdoor image classification," in *International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*. IEEE, 1998, pp. 42–51.

[4] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proceedings of ACM Multimedia Conference*, Berkeley, CA, November 2003, ACM.

[5] J.F.G. de Freitas P. Duygulu, K. Barnard and D.A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Seventh European Conf. on Computer Vision*, 2002.

[6] V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *Proceedings of ICASSP*, Montreal, Canada, May 2004, IEEE.

[7] NIST, *TREC Video Retrieval Evaluation Conference(TRECVID2003)*, Gaithersburg, MD, November 2003.

[8] A. Ratnaparkhi, *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1998.

[9] T. Utsuro, T. Miyata, and Y. Matsumoto, "Maximum entropy model learning of subcategorization preference," in *Proceedings of the 5th Workshop on VLC*, 1997, pp. 246–260.

[10] Jiwoon Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation," in *Image and Video Retrieval: Third International Conference, (CIVR)*, Dublin, Ireland, July 2004, Lecture Notes in Computer Science, Springer-Verlag.

[11] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, and et al, "IBM research TRECVID2003 video retrieval system," in *Proceedings of the TRECVID2003 Conference*. NIST, 2003.

[12] E. T. Jaynes, "Information theory and statistical mechanics," *The Physical Review*, vol. 106, pp. 620–630, 1957.

[13] Jason Baldridge, Tom Morton, and Gann Bierner, "openNLP maximum entropy modeling toolkit," http://maxent.sourceforge.net/, version 2.2.0, 2004.

[14] J.N. Darroch and D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.