USING REGIONAL INFORMATION IN LANGUAGE MODEL BASED AUTOMATIC CONCEPT ANNOTATION AND RETRIEVAL OF VIDEO

Dietrich Klakow

Spoken Language Systems
Saarland University
D-66125 Saarbruecken, Germany
Dietrich.Klakow@LSV.Uni-Saarland.De

ABSTRACT

In this paper we propose the use of regional information for the TREC Video Retrieval task of retrieving key-frames showing specific concepts in the image. We observe that the presence of a certain visual feature (e.g. color) is a strong indicator for a concept in particular if it occurs in a certain region of the image. We use this observation in a language model based image retrieval framework. This approach improves mean average precision from 0.187 (HMM based approach by Ghoshal et al.) to 0.221.

1. INTRODUCTION

The availability of large amounts of multimedia data make indexing and retrieval a relevant and challenging task. In the TREC Video Retrieval Evaluation [1] various systems are compared on a number of subtasks. In this paper we want to focus on the annotation of key-frames of broadcast news with concept labels like `text_overlay`, `outdoors` or `vehicle` and the retrieval based on these concepts. A very good description of the task can be found in [2].

There is a bulk of previous work. Duygulu et al. [3, 4] pioneered a machine translation based approach. The relevance model of Feng et al. [5] is based on a joint probability distribution of features and annotations. Last year, the JHU summer workshop provided a comparison of the two approaches and started the development of an GMM approach [6]. The latter was refined into an HMM and has shown good performance on the TREC-VID task [7]. Hence this approach will be used for comparison.

The core contribution of this work is the use of regional information, i.e. the observation that specific visual features are only a strong indicator for certain concept if the feature occurs in a specific region of the image. As a framework of annotation and retrieval we use language models. They are used in speech recognition already for a very long time and are a core component of any speech recognition system. For the retrieval of text based documents this approach has been proposed by [8] and was shown to outperform more traditional methods. Zhai and Lafferty [9] investigated various language model smoothing techniques for retrieval.

In this paper we want to first introduce a measure for the information available about a specific concept in a particular region of the image. In section 3 we formulate the retrieval models we use. Section 4 gives experimental results showing that certain regions of an image provide specific information about a concept and that this information is very important for image retrieval.

2. MEASURING THE IMPORTANCE OF REGIONAL INFORMATION

An important first step in designing a retrieval or classification task is an analysis of the dependencies of concepts (or classes) with the observations (or features) extracted from the data. Mutual information is an extremely useful tool for this purpose. For the specific task we want to measure how much information is contained in a certain region of an image and how much this tells about a specific concept. To this end we define a regional mutual information as

$$\mathrm{MI}(r,c) = \sum_v N(v,r,c) \log \left(\frac{N(v,r,c)}{N(v,r)N(c)} \right) \qquad . \qquad (1)$$

Here, c is a concept of the TREC-VID task. Typical examples for c are `face`, `text_overlay`, `weather_news` or `horse`. The regions are labeled with r and number rectangular regions in a 5x7 grid put on the image. In each region texture, color and edge information is extracted and quantized to a discrete set of so-called visterms V. An individual visterm is denoted by v in the formula. $N(v,r,c)$ counts how often a visterm v occurs in region r for an image labeled with concept c. $N(v,r)$ and $N(c)$ are the respective marginal frequencies, e.g. $N(c)$ counts how many images in the collection are annotated with concept c. We are averaging over all possible visterms because we want to know which region contains most information about the presence or absence of a specific concept. In the experimental section we will show that indeed very different patterns can be observed depending on the concept.

3. MODELS FOR RETRIEVAL

3.1. Language Models for Information Retrieval and Classification

Next we have to introduce a suitable retrieval framework. Ponte and Croft suggested in [8] language models to information retrieval from text collections and showed that they can outperform more traditional methods. We want to propose to use this technique also in image retrieval.

The key idea is to treat image retrieval as a classification problem. The Bayes classifier is defined by

$$\hat{c} = \text{argmax}_c P(v_1...v_R|c)P(c) \tag{2}$$

which provides minimum error rate if all probabilities are exactly known. Here, $P(v_1...v_R|c)$ is the probability that the set of visterms extracted from all regions 1 to R occur given a specific concept c. $P(c)$ is the prior for that concept. Unlike in previous work [9] we do not assume a uniform prior. The most likely concept according to (2) is assigned to an image. Alternatively a probability threshold can be set for assigning multiple concepts. We decided for a third variant that uses a separate binary classifier for each concept (concept present vs. concept absent).

Formula (2) has a data sparsity problem. We make the assumption that neighboring regions of an image are independent. Because of the very rough grid this is true to a very good degree. The established approaches used previously [6] assumed that the visterms are also independent of the region. This independence assumption would result in

$$P(v_1...v_R|c) = \prod_{r=1}^{R} P(v_r|c) \qquad . \tag{3}$$

We will use this model as a baseline. In contrast we propose to use the information about the region the visterm is coming from also in the classification process. This results in

$$P(v_1...v_R|c) = \prod_{r=1}^{R} P_r(v_r|c) \tag{4}$$

where $P_r(v_r|c)$ is the probability that a visterm occurs in a specific region given the concept.

Next we will describe how to estimate this probability. It is essential to avoid zero probabilities because that would exclude specific concepts from the retrieval. Hence language model smoothing methods come into play.

3.2. Absolute Discounting

Absolute discounting and its variants are the most popular smoothing techniques in speech recognition. It is defined by

$$P_r(v|c) = \frac{\max\left(N(v,r,c) - d, 0\right)}{N(r,c)} + \frac{dB}{N(r,c)} P_{BG}(v|c) \qquad . \tag{5}$$

where $N(v,r,c)$ is the frequency of joint observations of the visterm v together with concept c in region r and $N(r,c)$ the frequency of the concept in region r. $P_{BG}(v|c)$ is a background model used for smoothing with $P_{BG}(v|c)$ for all v and c. To obtain some probability mass for the background model the discounting parameter d reduces all seen events. B counts how often $N(v,r,c)$ is larger than d. The specific form of the pre-factor of $P_{BG}(v|c)$ can be derived from the normalization constraint of the overall model.

3.3. Dirichlet Prior

Using a Dirichlet prior results in

$$P_r(v|c) = \frac{N(v,r,c) + \mu P_{BG}(v|c)}{N(r,c) + \mu} \tag{6}$$

where μ is a smoothing parameter to be determined on the development data. Using a uniform background model results in so-called add-epsilon smoothing.

3.4. Linear Interpolation

Linear interpolation was first introduced by Jelinek and Mercer [10] and hence some people refer to it also as Jelinek-Mercer smoothing. It is defined by

$$P_r(v|c) = (1-\lambda)\frac{N(v,r,c)}{N(r,c)} + \lambda P_{BG}(v|c) \tag{7}$$

where $N(v,r,c)$ and $N(r,c)$ are frequencies on the training data, λ is a smoothing parameter to be tuned on the development data.

3.5. Background Models for Smoothing

As a background model we use two different versions. A uniform distribution - very often called a zerogram by people working in the area of language modeling - is defined by

$$P_{BG}^{Zero}(v|c) = \frac{1}{|V|} \tag{8}$$

where $|V|$ is the number of different visterms. A refined version is a unigram, that is the relative frequency, defined by

$$P_{BG}^{Uni}(v|c) = \frac{N(v)}{\sum_v N(v)} \qquad . \tag{9}$$

Note that both background models are independent of c. Both variants will be tested in the experimental section.

4. EXPERIMENTS

4.1. Data

We used the TREC-VID data as provided by IBM [11], [1]. It consists of key-frames from Broadcast News. Overall 75 different concept labels where used to annotate the images. Each image has on average 3.5 concept labels. On each image a 5x7 grid is used and color, texture and edge-strength is extracted as a 76-dimensional feature vector from each of the 35 regions. Later, vector quantization is used to produce 1000 different discrete visterms for color, texture and edge-strength.

The training data consists of 26195 images, the validation data used for tuning the parameters has 3955 images and the test data 9220 images. In [6] and [7] both, the continuous and the discrete version are used. There, the test set is referred to as `concept-fusion-2`.

4.2. Regional Mutual Information on TREC-VID

In Fig. 1 and Fig. 2 two examples of regional mutual information as defined in (1) are shown. For the concept `face`, (Fig. 1) most of the information can be found in the center of the image. This is quite intuitive as speakers on TV usually appear in the center of an image. Note that this figure accumulates information of many images and hence the size of the area with high mutual information does not allow to draw conclusions about the size of a face. The second example (Fig. 2) is for the concept `text_overlay`. Here, most information can be found is the lower part of the images. This is the example with the most strongly localized area providing information about a concept. Interesting is also the concept `weather_news` not shown here due to space limitations. Here, most information comes from the corners because they have a specific background typical for the broadcasts recorded for TREC-VID. Some other examples (e.g. `vehicle`) are hard to explain.

4.3. Retrieval Experiments

Finally, we want to turn to retrieval experiments. The queries are the specific concepts. Table 1 summarizes our results.

We compare the three different smoothing techniques using zerogram and unigram background models. All six models are tested without using regional information (baseline) and with regional information.

If regional information is not modeled, the specific smoothing method and background model is not important because there is hardly any data sparsity problem. This changes if regional information is modeled as well. This boosts performance from a mean average precision of 0.150 up to 0.221 in the best case. Now smoothing is important. In general using a unigram background model is better than zerogram background model which is intuitive as the unigram is more specific. Among the three smoothing methods all three are competitive but linear interpolation outperforms the other two slightly which is consistent with previous findings from text based retrieval using language models.

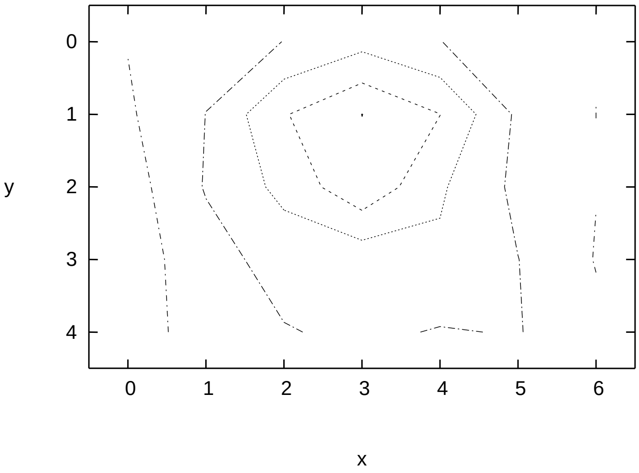

Fig. 1. Regional mutual information for the concept ”face”. x and y are the coordinates of the regions in the image.

	Baseline		+ regional info.	
Model	BG: Zero	BG: Unigram	BG: Zero	BG: Unigram
Absolute disc.	0.149	0.149	0.209	0.215
Dirichlet prior	0.149	0.150	0.207	0.218
Linear interp.	0.145	0.148	0.215	0.221

Table 1. Comparison of the mean average precision (mAP) of different retrieval models without and with using the regional information. Note that the best known result on this data set so far is 0.187 [7].

In Fig. 3 we show the precision-recall graph. We compare all three smoothing method using a unigram background model and regional information with the HMM approach. Linear interpolation is again slightly better than the other two language models, which show almost identical performance. All three approaches outperform the HMM.

5. CONCLUSION AND FUTURE WORK

In this paper we have investigated the use of regional information in concept based video retrieval using a language model based retrieval framework. In particular including the region from which visterms are extracted in the modeling is very essential. Also picking the right smoothing method (in this case linear interpolation with a unigram as a background model) helps. We have shown that this method outperforms previous

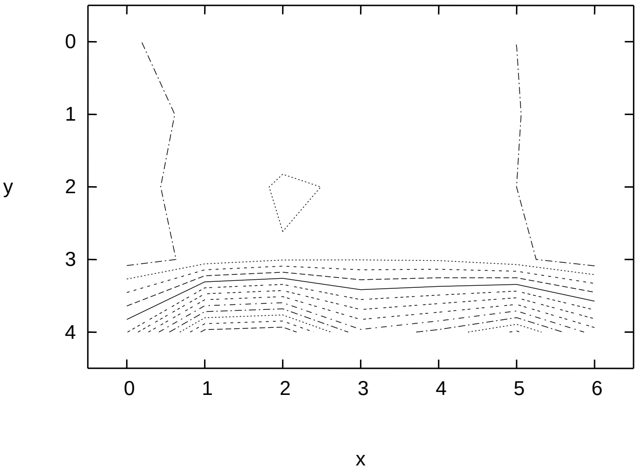

Fig. 2. Regional mutual information for the concept ”text_overlay”.

approaches by improving mAP from 0.187 to 0.221.

However, this investigation can only be a very first step because there are clear opportunities for future directions of research. On the one hand side, the way regional information is modeled is still very simple. More refined models could for example use tying among neighboring regions to make the models more robust. Also, the speech community has developed are large set of different language model types that can be used for future experiments.

6. ACKNOWLEDGMENT

The author would like to thank G. Iyengar, P. Duygulu, S. P. Khudanpur, M. R. Krause, R. Manmatha and H. J. Nock for many interesting discussions. Particular thanks goes to IBM (Giri Iyengar) for providing the data.

7. REFERENCES

[1] *TREC Video Retrieval Evaluation Online Proceedings*, vol. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2004.

[2] Wessel Kraaij, Alan F. Smeaton, Paul Over, and Joaquim Arlandis, “Trecvid 2004 - an overview,” in *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[3] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” in *Proc. European Conference on Computer Vision*, 2002.

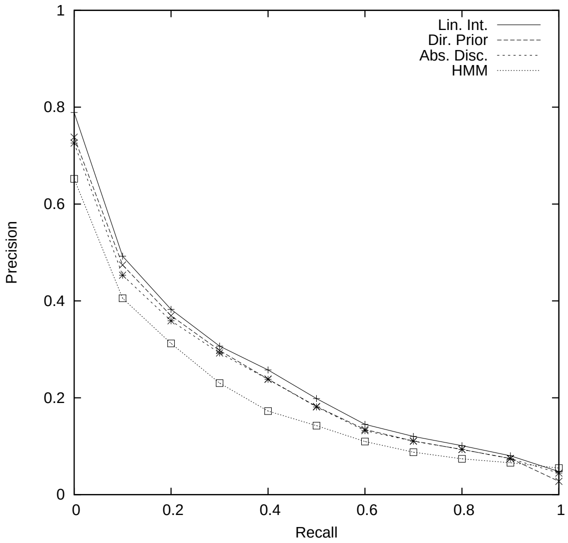

Fig. 3. Precision recall curve for the proposed method compared with a recently published HMM approach [7].

[4] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan, “Matching words and pictures,” *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[5] S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *IEEE CVPR*, 2004.

[6] Giri Iyengar et al., “Joint visualtext modeling for automatic retrieval of multimedia documents,” in *ACM Multimedia*, 2005.

[7] Arnab Ghoshal and Sanjeev Khudanpur, “Modeling word co-occurrence for automatic image annotation and retrieval,” in *Proc. SIGIR*, 2005.

[8] Jay M. Ponte and W. Bruce Croft, “A language modeling approach to information retrieval,” in *Proc. SIGIR*, 1998.

[9] Chengxiang Zhai and John Lafferty, “A study of smoothing methods for language models applied to ad hoc information retrieval,” in *Prof. SIGIR*, 2001.

[10] Hermann Ney, Ute Essen, and Reinhard Kneser, “On structuring probabilistic dependencies in stochastic language modeling,” *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.

[11] A. Amir et al., “Ibm research trecvid-2004 video retrieval system,” in *Proc. NIST Text Retrieval Conf. (TREC)*, 2004.