

SOURCE ADAPTATION FOR IMPROVED CONTENT-BASED VIDEO RETRIEVAL

Arnab Ghoshal and Sanjeev Khudanpur

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Adaptation of hidden Markov model (HMM) parameters to individual speakers is known to provide considerable improvements over speaker-independent speech recognition systems. This paper applies this idea of model adaptation to a content-based video retrieval system that uses HMMs, with different *sources* of video treated analogously to different speakers. Source-independent HMMs are adapted to each video-source using the maximum a posteriori probability (MAP) and maximum likelihood linear regression (MLLR) techniques. It is shown that MLLR is not effective in modeling source variability in video, while MAP is highly effective. An overall improvement of 39% is demonstrated in video retrieval performance on the TRECVID 2005 benchmark test over a competitive baseline system via source-adaptation and improved use of the HMM likelihoods in retrieval.

1. INTRODUCTION

The content of communications in the digital age is increasingly multi-modal in nature, with text, images and even speech or video being used in a single “document.” Content-based indexing and retrieval of multimedia is therefore becoming an increasingly important issue. Unlike text retrieval, where the modality in which the user usually specifies her information need is the same as the modality of the search collection, there is relatively little work in image and video retrieval based on textual queries. Important progress has been made in the last few years in content-based image retrieval, as reported by Duygulu et al [1], Blei et al [2], Jeon et al [3] and others.

While the classical image understanding problem, *i.e.* the problem of recognizing all the objects in a given image, is very difficult due to several invariance issues, an aspect of the image and video indexing and retrieval problem that makes it relatively more tractable is the availability of *side information*: images in multimedia documents are often accompanied by descriptive text that a model may use to infer the content of an image, and video is often accompanied by speech. With this consideration, we [4] have recently developed a joint stochastic model, specifically a hidden Markov model (HMM), for images and their accompanying captions. HMM parameters are estimated from a manually annotated (training) collection of image+caption pairs; the caption-words are from a large but fixed vocabulary of objects or concepts.

In this paper, we report two *significant* improvements to the model of [4]. In particular, we study the adaptation of HMM parameters to different video sources, and a *concept-specific* variation of the model in which, for each concept in the vocabulary, we train one HMM on all the images containing the concept and another HMM on all images not containing the concept. For each test image, presence or absence of a concept is determined by a *likelihood ratio test* under these two models. We also study the adaptation of the concept-specific models to individual video sources.

This paper is organized as follows. Section 2 formally describes the two types of HMMs used for image annotation. Section 3 describes HMM adaptation techniques. Section 4 presents a series of experimental results on the NIST TRECVID 2005 data-set, followed by a discussion in Section 5.

2. HMMS FOR IMAGE ANNOTATION

Let a collection $\mathcal{L} \equiv \{(I, C)\}$ of image+caption pairs be given. Let $I \equiv \{i_1, \dots, i_T\}$ denote image-segments (image-regions), and $C \equiv \{c_1, \dots, c_N\}$ the objects (concepts) present in that image, as specified by the label (caption). The T image-regions may be object-based, with each region corresponding to one semantically distinct object, or they may be a simple rectangular partition of the image into fixed-size blocks. For each image-region i_t , $t = 1, \dots, T$, let $x_t \in \mathbb{R}^d$ represent color, texture, edges, shape and other salient visual features of the region. Let \mathcal{V} denote the total vocabulary of the caption-words c_n across the entire collection of images.

We propose to model the visual features $\{x_1, \dots, x_T\}$ as a hidden Markov process, generated by an unobserved underlying Markov chain $\{s_t\}$ with a known initial state s_0 and transition probabilities $p(s_t|s_{t-1})$. We model the output density for each state s as a mixture of multivariate Gaussian densities on \mathbb{R}^d :

$$f(x|s) = \sum_{m=1}^M w_{m,s} \frac{e^{-\frac{1}{2}(x-\mu_{m,s})^T \Sigma_{m,s}^{-1}(x-\mu_{m,s})}}{\sqrt{(2\pi)^d |\Sigma_{m,s}|}}, \quad (1)$$

where $w_{m,s}$ is the mixture weight, $\mu_{m,s}$ the mean-vector and $\Sigma_{m,s}$ the diagonal covariance-matrix of the m -th mixture component of state s .

The joint likelihood of a state sequence $s_1^T \equiv \{s_1, \dots, s_T\}$ and features $x_1^T \equiv \{x_1, \dots, x_T\}$ is

$$f(x_1^T, s_1^T | s_0) = \prod_{t=1}^T f(x_t | s_t) p(s_t | s_{t-1}). \quad (2)$$

The model proposed in [4] associates one state s with each word in the concept vocabulary \mathcal{V} , as summarized in Section 2.1, formalizing the notion that each image region is a stochastic realization of one of the concepts present in the image. We propose an alternative model in Section 2.2, where the states have no such semantic interpretation, but instead model spatial locality of the visual features.

2.1. A Joint Model of Visual Features and Captions

In the joint model, the states $\{s_t\}$ of the underlying Markov chain for an image I take values in C , its caption. A label (or concept) $c \in \mathcal{V}$ appearing in two different images is modeled by the same state, and the HMMs for all images “share” states from a common pool of $|\mathcal{V}|$

tioned states. For an image+caption pair (I, C) , $s_1^T \equiv \{s_1, \dots, s_T\} \in C^T$, with $C \subset \mathcal{V}$.

Note that knowing the state sequence $\{s_t\}$ is equivalent to having the *alignment* of each image-region i_t with one of the words in the caption. Even though this level of detail is generally not provided in captions, an HMM readily provides the joint likelihood of an image+caption pair $(I, C) \equiv (x_1^T, C)$ via the forward-algorithm.

$$f(x_1^T, C|s_0) = \sum_{s_1^T \in C^T} \prod_{t=1}^T f(x_t|s_t) p(s_t|s_{t-1}). \quad (3)$$

Furthermore, given a training collection of image+caption pairs, *emission* densities $f(x|c)$ and *transition* probabilities $p(c'|c)$ of the HMM may be estimated to maximize the likelihood (3) of the training pairs. Details of this maximum likelihood estimation procedure are standard and therefore omitted (cf [5]).

For indexing a new image I , the HMM provides the conditional probability, given all the visual evidence x_1^T in I , that an image-region i_t was generated by a concept $c \in \mathcal{V}$, as

$$\begin{aligned} p(s_t = c|x_1^T, s_0) &= \frac{f(x_1^T, s_t = c|s_0)}{f(x_1^T|s_0)} \\ &= \frac{\sum_{s_1^T: s_t=c} \prod_{t=1}^T f(x_t|s_t) p(s_t|s_{t-1})}{\sum_{s_1^T \in \mathcal{V}^T} \prod_{t=1}^T f(x_t|s_t) p(s_t|s_{t-1})}. \end{aligned} \quad (4)$$

Therefore, the probability of a particular concept $c \in \mathcal{V}$ being present (somewhere) in an image may be calculated as

$$p(c|I, s_0) = \frac{1}{T} \sum_{t=1}^T p(s_t = c|x_1^T, s_0). \quad (5)$$

Unlabeled images in a test collection $\{I\}$ may therefore be ranked for the presence of any particular concept c based on this posterior probability. In other words, the relevance score assigned to an image I for a query c is

$$\text{score}(I, c) = p(c|I, s_0). \quad (6)$$

See [4] for details of this model and its retrieval performance.

2.2. Concept-Specific Image Models

Image retrieval for a text query c is essentially the task of deciding, for each image I , whether or not it contains the concept c . This motivates the following concept-specific model, in which one pair of HMMs is estimated for each concept: an HMM \mathcal{H}_c^+ from all images that contain the concept c , and an \mathcal{H}_c^- from all images that do not contain the concept. Both \mathcal{H}_c^+ and \mathcal{H}_c^- have simple left-to-right topologies, with as many states as there are rectangular image regions, as shown in Figure 1. Formally, the state space of both \mathcal{H}_c^+ and \mathcal{H}_c^- is $\mathcal{S} = \{1, \dots, T\}$, and unlike the preceding joint model, these states do not have any interpretation in terms of the concept vocabulary \mathcal{V} and instead simply model spatial properties of the images.

With \mathcal{H} generically denoting either \mathcal{H}_c^+ or \mathcal{H}_c^- for some concept c , the marginal likelihood of an image I under \mathcal{H} is

$$f(I|\mathcal{H}) \equiv f(x_1^T|\mathcal{H}, s_0) = \sum_{s_1^T \in \mathcal{S}^T} \prod_{t=1}^T f_{\mathcal{H}}(x_t|s_t) p_{\mathcal{H}}(s_t|s_{t-1}).$$

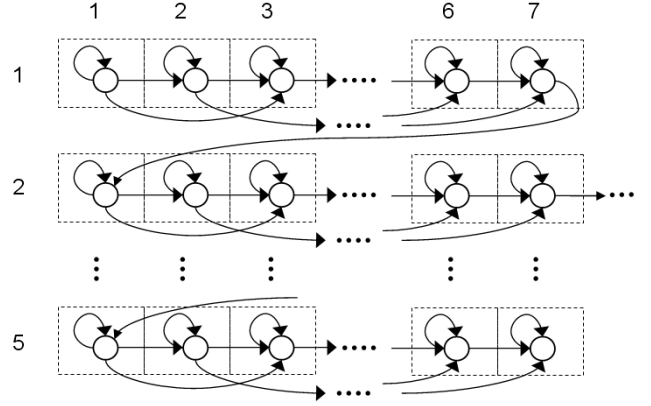


Fig. 1. HMM Topology for the Concept-Specific Models.

The parameters $f_{\mathcal{H}_c^+}(\cdot|s)$ and $p_{\mathcal{H}_c^+}(s'|s)$ of each \mathcal{H}_c^+ are chosen to maximize the likelihood of images containing c :

$$\mathcal{H}_c^+ = \arg \max_{\mathcal{H}} \prod_{I_l: c \in C_l} f(I_l|\mathcal{H}), \quad (7)$$

where $\mathcal{L} \equiv \{(I_1, C_1), \dots, (I_L, C_L)\}$ denotes the training set. Similarly the parameters of each \mathcal{H}_c^- are estimated to maximize the likelihood of images not containing c . Note that this is just maximum likelihood estimation, albeit carried out using a different partition of \mathcal{L} for every pair of HMMs in the family $\{\mathcal{H}_c^+, \mathcal{H}_c^-, c \in \mathcal{V}\}$.

For an unlabeled image collection $\{I\}$, we calculate the likelihoods $f(I|\mathcal{H}_c^+)$ and $f(I|\mathcal{H}_c^-)$ for each concept c , and then rank-order the images according to the likelihood ratio

$$\text{score}(I, c) = \frac{f(I|\mathcal{H}_c^+)}{f(I|\mathcal{H}_c^-)} = \frac{f(x_1^T|\mathcal{H}_c^+, s_0)}{f(x_1^T|\mathcal{H}_c^-, s_0)}. \quad (8)$$

In practice, we have found that the Viterbi approximation

$$\text{score}(I, c) \approx \frac{\max_{s_1^T} \prod_{t=1}^T f_{\mathcal{H}_c^+}(x_t|s_t) p_{\mathcal{H}_c^+}(s_t|s_{t-1})}{\max_{s_1^T} \prod_{t=1}^T f_{\mathcal{H}_c^-}(x_t|s_t) p_{\mathcal{H}_c^-}(s_t|s_{t-1})}$$

results in nearly identical retrieval performance.

3. ADAPTING TO INDIVIDUAL VIDEO SOURCES

Speaker-dependent (SD) automatic speech recognition systems perform much better than speaker-independent (SI) systems [6]. However, training an SD system is often impractical, since it requires obtaining a large amount of speaker specific training data. Speaker adaptation provides a convenient way to “tune” a SI system to individual speakers with comparatively little data from each speaker. Model adaptation techniques, where the parameters of a SI system are adjusted to better match individual speakers, may be broadly classified into Bayesian approaches [6, 7] and transformation based approaches [8, 9]. The adaptation is said to be supervised or unsupervised, depending on whether a manual transcription of the speech is provided or not.

Inspired by these results, we investigate supervised adaptation for HMM-based image retrieval. We treat each video-source as a different speaker, and adapt a source-independent (SI) HMM system to perform source-dependent (SD) retrieval.

3.1. Adaptive Transforms (MLLR)

The maximum likelihood linear regression (MLLR) approach adapts the parameters of an SI system by applying an affine transform [9]. Often, only the Gaussian means of the SI system are transformed: the mean vectors $\mu_{m,s}$ of each mixture component m of every state s are replaced, respectively, with

$$\hat{\mu}_{m,s}^{(k)} = \mathbf{A}^{(k)} \mu_{m,s} + b^{(k)}, \quad (9)$$

where the SD transform-parameters $[\mathbf{A}^{(k)}, b^{(k)}]$ are estimated separately for each video source $k = 1, \dots, K$, to maximize the likelihood of the source-specific training data. The estimation procedure is well known [8].

3.2. Bayesian Adaptation (MAP)

A Bayesian approach to source-adaptation is to consider the SD mean vectors $\mu_{m,s}^{(k)}$ for a source k to themselves be random vectors and, given some source-specific training data, to compute the values of the SD means with the maximum *a posteriori* probability (MAP) [6]. In particular, if we assume *a priori* that $\mu_{m,s}^{(k)}$ is Gaussian with mean equal to its SI value $\hat{\mu}_{m,s}$, and variance τ^{-2} , then the MAP estimate of $\mu_{m,s}^{(k)}$ given some source-specific data is

$$\hat{\mu}_{m,s}^{(k)} = \frac{\sum_{t=1}^T \gamma_{m,s}(t) x_t^{(k)} + \tau \hat{\mu}_{m,s}}{\sum_{t=1}^T \gamma_{m,s}(t) + \tau}, \quad (10)$$

where $\gamma_{m,s}(t)$ is the posterior probability under the SI HMM that $x_t^{(k)}$, the t -th image feature of video-source k , was “emitted” by mixture component m of state s . Setting $\tau = 0$ results in maximum likelihood estimation of the SD means from only the source-specific data, which may result in over-fitting. τ is chosen empirically to temper this effect [7].

4. EXPERIMENTAL RESULTS

The NIST TRECVID 2005 benchmark test provides an evaluation forum for video retrieval from text queries. About 170 hours of broadcast quality news videos from 13 different programs in English, Arabic and Mandarin are provided, along with shot boundaries. For a detailed description of the TRECVID 2005 data-set refer to: <http://www-nlpir.nist.gov/projects/tv2005/tv2005.html>. The data-set also provides one or more keyframes for each shot, resulting in about 152K keyframes. These are divided into a 74K-keyframe development (DEV) set and 78K-keyframe evaluation (EVAL) set. The *high-level feature detection* task is to detect the presence or absence of 10 predetermined *benchmark* concepts (high-level features) in each shot of the EVAL partition. Using each of the 10 concepts as single-word queries, systems are required to return ranked-lists of up to 2000 shots, and system performance is measured via non-interpolated *mean average precision* (mAP), a standard metric for document retrieval. For more information, see the TREC-10 Proceedings appendix on common evaluation measures, available at <http://trec.nist.gov/pubs/trec10/appendices/measures.pdf>.

Each keyframe in DEV is manually marked for the presence or absence of each of the 10 benchmark concepts, as well as 29 other concepts. For each keyframe I , we create a caption C by noting the concepts present in it. Therefore, $|\mathcal{V}| = 39$, and since many keyframes do not contain any of the 39 concepts, $C = \phi$ for about 29% of the (I, C) pairs. We further divide DEV into a training

(TRN) set \mathcal{L} of 57K keyframe+caption pairs, and a check (CHK) set of 17K.

Though the high-level feature detection task permits analysis of entire shots before retrieval, the techniques described in Section 2 currently handle only still images. In particular, the scores of (6) and (8) are used to rank-order the keyframes of CHK or EVAL in response to queries c . Since the task requires ranking entire shots in response to each query, and some shots contain multiple keyframes, we derive a ranked-list of shots from the ranked-list of keyframes by a simple scheme — the rank of a shot is the harmonic mean of the ranks of its keyframes. This scheme could clearly be improved upon, but it does not affect the main results of this paper.

Visual features x_1^T were extracted for each keyframe using a 5×7 rectangular partition, and provided to us by Giridharan Iyengar of IBM. The 80-dim features capture color moments, oriented-edges, and texture in each sub-image [10].

4.1. Adaptation Results for the Baseline HMM System

We first train the HMM of Section 2.1 on all (I, C) pairs with $C \neq \phi$ from all 13 video sources in TRN. For each of the 39 concepts c , we then rank *all* the keyframes I in CHK according to (6). This HMM, with $M = 100$ Gaussian densities per mixture, was shown in [4] to be comparable to the state of the art on the TRECVID 2003 benchmark test, and the mAP of this system over 39 concepts on CHK forms the *baseline* for subsequent experiments. The mAP of this system on EVAL for the 10 benchmark concepts is also the baseline for the final comparison in Section 5.

We next perform 3-4 iterations of supervised adaptation of the SI HMM, either MLLR or MAP, for each of the 13 sources, and again measure video retrieval performance on CHK. The resulting mAP for all 39 concepts, and for the subset of 10 benchmark concepts, is reported in Table 1.

System	Baseline	MLLR	MAP
mAP (39)	0.230	0.231	0.243
p-value	—	0.46	0.0001
mAP (10)	0.185	0.186	0.196
p-value	—	0.42	0.0213

Table 1. Improvements in mAP of ranked-retrieval of *all* shots in the CHK set, and their statistical significance, due to source-adaptation of the HMMs of [4].

Note that significant improvements are obtained by MAP adaptation, but almost none by MLLR. We conclude that MLLR, while effective for channel compensation in speech recognition, is not appropriate for capturing source-specific image-variability.

4.2. Adapting the Concept-Specific Image Models

We next train the HMMs of Section 2.2 on all (I, C) pairs with $C \neq \phi$ from all 13 video sources in TRN. The HMM \mathcal{H}_c^+ is trained on images from TRN containing the concept c , and \mathcal{H}_c^- on images not containing c . We study two alternatives for *initializing* the iterative estimation of the HMMs. One is to follow the usual “flat-start” recipe (cf [11]), and the other is to first fully train up a *background* HMM \mathcal{H} using all the images in TRN, and use it to initialize \mathcal{H}_c^+ and \mathcal{H}_c^- . Since there are fewer training images, particularly for \mathcal{H}_c^+ , the HMMs in this section have $M = 10$ to 20 Gaussian densities per mixture.

We then perform 3-4 iterations of supervised adaptation of the SI HMMs, either MLLR or MAP, for each of the 13 sources, and again measure video retrieval performance on CHK. The resulting mAP is reported in Table 2.

Initialization System	Flat-start			Background	
	SI	MLLR	MAP	SI	MAP
mAP (39)	0.282	0.284	0.295	0.296	0.304
p-value	—	0.19	0.0001	—	0.001
mAP (10)	0.209	0.214	0.228	0.227	0.242
p-value	—	0.24	0.01	—	0.01

Table 2. The mAP of ranked-retrieval of *all* shots in the CHK set following source adaptation of concept-specific HMMs.

Note, first, that the concept-specific HMMs significantly outperform the HMM of Section 4.1. Furthermore, the trend of MLLR not being beneficial and MAP providing further significant gains continues to hold for the “flat-start” training of HMMs. For this reason, we did not perform MLLR for the “background” initialized HMMs. The MAP adapted, “background”-initialized HMMs is dramatically better than the baseline. The mAP of 0.242 is also significantly better than the system of [10], which attains an mAP of 0.197 on this training-and-test configuration.

4.3. Results on the TRECVID 2005 Task

To further confirm these advances, we evaluated the baseline system of Section 4.1 and the best model of Table 2 on the EVAL partition, and the mAP for the 10 benchmark concepts in Table 3 indicates an overall improvement of 39% over the HMM system of [4].

System	Baseline	“background”+MAP
mAP (10)	0.137	0.225

Table 3. Performance of the baseline of Section 4.1 and the best of Section 4.2 measured on the top 2000 retrieved shots from the TRECVID 2005 benchmark test.

We also remark that the system [10], trained on the entire DEV data obtains an mAP of 0.198 on the TRECVID 2005 task, while the “background”+MAP system obtains a significantly higher mAP even when trained on only the TRN portion of DEV.

5. CONCLUDING REMARKS

We have demonstrated remarkable improvement in retrieval of video from text queries using HMMs within the context of the TRECVID 2005 task. In particular, we have shown that both source-adaptation and use of likelihood ratios instead of likelihoods in ranking frames or shots yield significant improvements in video retrieval performance. This coupled with the computational efficiency of HMMs makes them highly suitable for large-scale image and video retrieval. We are investigating discriminative techniques for estimating the joint video-text HMM so that the benefits of joint modeling as well as likelihood-ratio based ranking are obtained in a unified model.

6. ACKNOWLEDGMENTS

We gratefully acknowledge Giridharan Iyengar, who assisted significantly in this research by providing us with the visual features for the experiments, and Janne Argillander, for providing us with the IBM system’s performance on the training and test partitions used here. We also thank Brock Pytlik for many helpful discussions and for his efforts in carrying out the TRECVID evaluations, where he has explored many other novel models. This work was partially supported by NSF Grant No IIS-0121285 and DoD Contract No MDA904-01-C-1005.

7. REFERENCES

- [1] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, “Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary,” in *Proc. ECCV*, 2002, vol. 4, pp. 97–112.
- [2] D. M. Blei and M. I. Jordan, “Modeling Annotated Data,” in *Proc. ACM SIGIR*, 2003, pp. 127–134.
- [3] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic Image Annotation and Retrieval using Cross-Media Relevance Models,” in *Proc. ACM SIGIR*, 2003, pp. 119–126.
- [4] A. Ghoshal, P. Ircing, and S. Khudanpur, “Hidden Markov Models for Automatic Annotation and Content-Based Retrieval of Images and Video,” in *Proc. ACM SIGIR*, 2005, pp. 544–551.
- [5] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [6] C-H Lee, C-H Lin, and B-H Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 806–814, April 1991.
- [7] J-L Gauvain and C-H Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.
- [8] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.
- [9] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, “Speaker adaptation using constrained estimation of gaussian mixtures,” *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, September 1995.
- [10] J. Argillander, G. Iyengar, and H. J. Nock, “Semantic Annotation of Multimedia using Maximum Entropy Models,” in *Proc. ICASSP*, March 2005.
- [11] S. Young et al, *The HTK Book*, 2002.