RC23647 (W0506-140) June 30, 2005
Computer Science

IBM Research Report

Autonomous Learning of Visual Concept Models

Xiaodan Song*, Ching-Yung Lin, Ming-Ting Sun*
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

*Department of Electrical Engineering
University of Washington
Box 352500
Seattle, WA 98195

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Autonomous Learning of Visual Concept Models

Xiaodan Song[1], Ching-Yung Lin[2], and Ming-Ting Sun[1]

[1]: Dept. of Electrical Engineering, University of Washington, Box 352500, Seattle, WA 98195
[2]: IBM T. J. Watson Research Center, Hawthorne, NY 10532
song@ee.washington.edu, chingyung@us.ibm.com, sun@ee.washington.edu

***Abstract*—As the amount of video data increases, organizing and retrieving video data based on their semantics is becoming more and more important. Traditionally, supervised learning is used to build models for detecting semantic concepts. However, in order to obtain a substantial amount of training data, extensive labeling work is needed with the supervised learning schemes. In this paper, we propose a novel Autonomous Learning mechanism, in which imperfect information extracted from cross-modality information is used for training. This shall, thus, not only reduce the number of examples needed for labeling, as active learning and transductive learning do, but also totally avoid the manual labeling process. First of all, imperfect labels without user involvement are obtained from cross-modality information. Then based on our proposed new schemes, "Generalized Multiple-Instance Learning" and "Uncertain Labeling Density", the system conjectures relevance score of visual concepts. From these scores, Support Vector Regression is used to build generic visual models. Our proposed algorithm is tested on several concepts in large video databases. Preliminary experiments show promising results in limited number of concepts. This novel Autonomous Learning mechanism can achieve better system average precisions than two supervised algorithms. Currently, we're investigating the performance of this proposed system by taking large scale experiments, whose results is not in this paper yet.**

I. INTRODUCTION

As the amount of broadcast video data increases, content-based video indexing and retrieval is becoming increasingly important. Supervised machine learning methods has shown its effectiveness on modeling generic visual models [10]. Although it has the best performance on the NIST TREC concept detection benchmarking (2002-2004), a huge amount of work is required to manually label each image in large video datasets, and any new concepts not previously labeled would not be able to be dealt with. With this extensive labeling in sight, many methodologies were proposed to reduce the number of labeled instances, e.g., active learning and transductive learning [1][2]. In pool-based active learning [1], the learner has access to a pool of unlabeled instances and can request the labels for some of them. It is hoped that allowing the learner this flexibility will reduce the learner's need for large quantities of labeled data. In transductive learning [2], both the labeled and unlabeled samples are exploited into learning process. However, both of them still need user inputs. Another difficulty for generic visual concept learning is that when the size of positive examples in the training data is small, the performance of supervised learning algorithms could be affected significantly, especially when the positive examples are not so informative for learning the desired concept. The traditional solution to this problem is the relevance feedback technique [3]. During the retrieval process, the user interactively selects the most relevant images and provides a weight according to the preference for each relevant image. By dynamically updating weights based on the feedback, user's high level query and perception subjectivity is captured. However, this relevance feedback technique increases the burden of the users even more. In many practical applications, it is desirable if we can have an automatic algorithm, which does not need costly supervision and relevance weighting process.

Cross-modality data provide possibilities for the automatic learning mechanism. In [4], we used the association between the content in visual data and audio data in video sequences to develop an automatic training scheme for face recognition from large video databases and get promising results. Later in [5], we tried to use the correlation between the textual and the visual modalities for the image data available on the web to build models for content-based image retrieval. In this paper we generalize the ideas to a learning mechanism – Autonomous Learning mechanism and use this mechanism to obtain generic visual models from video sequences. The overall process of the Autonomous Learning mechanism includes three steps in total: imperfect labeling, uncertainty pruning, and relevance modeling (Fig. 1). In the step of "imperfect labeling", the correlation between the data of different modalities are used to get imperfect labels. In this paper, the imperfect labels, are obtained by the association between audio and visual data. Then in the step of "Uncertainty Pruning", a relevance rank list is achieved by our proposed "Generalized Multiple-Instance Learning" (GMIL) and "Uncertain Labeling Density" (ULD) algorithm, which is called as GMIL-ULD for convenience, from the imperfect labels. Finally in the step of "Relevance Modeling", Support Vector Regression (SVR) is used to build the generic visual models from the relevance ranking list so that the most informative examples are used to boost the retrieval performance. Fig. 2 illustrates the data involved in different steps.

The rest of the paper is organized as follows. In Section 2, we propose the Autonomous Learning mechanism in details. In Section 3, experimental results are presented. We show conclusions and future directions in Sections 4.

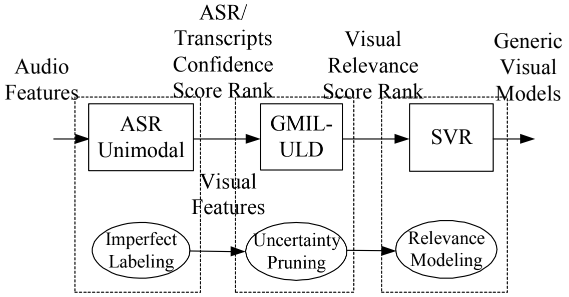
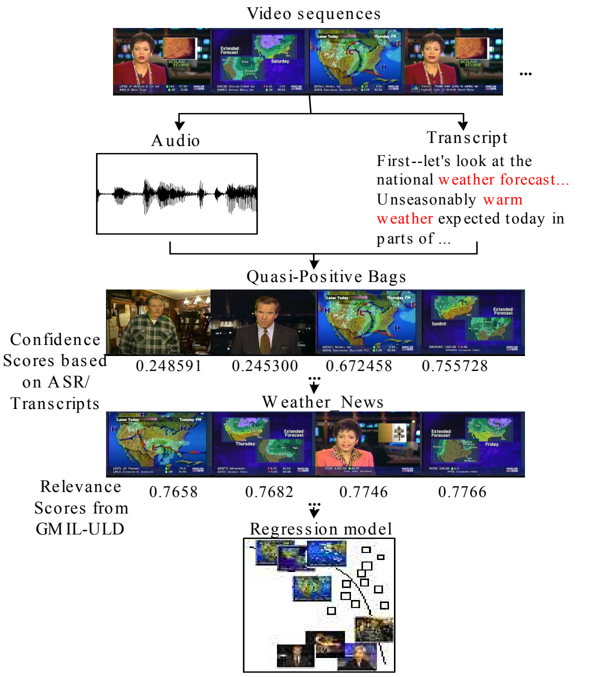


Figure 1. Block diagram of Autonomous Learning



First--let's look at the national weather forecast... Unseasonably warm weather expected today in parts of ...

Figure 2. Autonomous Learning based on cross-modality information

II. AUTONOMOUS LEARNING

In this section, we introduce two novel schemes: "Generalized Multiple Instance Learning" and "Quasi-Positive bag", and "Uncertain Labeling Density". Then, we describe the "Relevance Learning" scheme.

A. *GMIL and Quasi-Positive Bag*

Multiple Instance Learning (MIL) was recently proposed for machine learning to solve the ambiguity in the manual labeling process by making weaker assumptions about the labeling information [6][7][8]. In this learning scheme, instead of giving the learner labels for individual examples, the trainer only labels collections of examples, which are called bags. A bag is labeled negative if all the examples in it are negative, and labeled positive if there is at least one positive example in it. The key challenge in MIL is to cope with the ambiguity of not knowing which instances in a positive bag are actually positive and which are not. MIL helps to deal with the ambiguity in the manual labeling process. However, users still have to label the bags in the MIL framework. To prevent the tedious manual labeling work, we need to generate positive bags and negative bags automatically. In practical applications, it is very difficult if not impossible to automatically generate the positive and negative bags reliably.

In our scenario, although there is a relatively high probability that the concept of interest (e.g. a person's face) may appear in the visual data when related features (e.g. the person's name) appear in the audio data, there are many cases where no such an association exists. For example, in a talk show, a person's name may be mentioned many times without any pictures of that person's face. To overcome this problem, we extend the concept of "positive bags" to "Quasi-Positive bags". A "Quasi-Positive bag" has a high probability to contain a positive instance, but is not guaranteed to contain one. With the introduction of "Quasi-Positive bags", "Generalized Multiple-Instance Learning" is proposed to removes a major limitation of applying MIL to many practical problems. In GMIL, a bag is labeled negative, if all the instances in it are negative. A bag is Quasi-Positive, if in a high probability, at least one instance in it is positive. We extract the Quasi-Positive bags based on the hypothesis that the features in the audio data are highly correlated with the concepts in the visual data.

B. *Uncertain Labeling Density*

Among the algorithms proposed to solve MIL [6][7][8], Diversity Density (DD) algorithm [6], which is a measure of the intersection of the positive bags minus the union of the negative bags, is not highly dependent on the distribution of the negative bags. It can even solve the problem without the existence of any negative bags. By finding the maximal DD, we can find the desired concept as the feature vector that is the intersection of instances in positive bags, and not in negative bags. In our application, what we have are Quasi-Positive bags, i.e., unreliable positive bags that have high probabilities but are not guaranteed to contain positive instances, i.e., there are false-positive bags. In a false-positive bag, by the original DD definition, $\Pr\left(t \mid B_i^+\right)$ is very small or even zero. These outliers will influence the DD significantly due to the multiplication of the probabilities defined in the original DD formulation [6].

To avoid the influence of false-positive bags, we propose "Uncertain Labeling Density" (ULD) to handle the Quasi-Positive bag problem. The ULD is defined as:

$$\arg\max_t \frac{\sum_i \Pr\left(t \mid B_i^+\right) \bullet \prod_i \Pr\left(t \mid B_i^-\right)}{Z} \tag{1}$$

where $\Pr\left(t \mid B_i^+\right)$ and $\Pr\left(t \mid B_i^-\right)$ are also estimated by the noise-or model [8], and Z is a normalization parameter. By using summation instead of multiplication, the false-positive bags which possess smaller $\Pr\left(t \mid B_i^+\right)$ will make little influence to the ULD value, when t is the true concept, while the truly positive bags will still contribute large values to the ULD.

C. Relevance Learning

Based on the ASR unimodal analysis results [9], each shot will be associated with a confidence score, in the range of $[0,1]$, showing how likely this shot belongs to this concept from the view point of audio features. Based on these scores, we choose the shots with nonzero confidence scores as the Quasi-Positive bags. In this paper, we do not use negative bags when calculating ULD because ASR based analysis is not so accurate to tell which examples are definitely unrelated.

Considering the reliability of each positive bag, $\Pr\left(t \mid B_i^+\right)$ is calculated as:

$$\Pr\left(t \mid B_i^+\right) = 1 - \prod_j \left[\left(1 - \Pr\left(t \mid B_{ij}^+\right)\right) \bullet CS(i)\right] \tag{2}$$

where $CS(i)$ represents the confidence score for the ith shot. The more reliable the positive bag, the more contribution to the whole density it provides.

Based on Quasi-Positive bags and the GMIL-ULD algorithm, the point with the highest ULD value is chosen as the visual model for the concept we are trying to learn, denoted as x_E. Then, the visual rank list is generated by considering both the distances between the instances and the learned most informative example, and the ULD values:

$$CS_v(i) = EDD(i) \cdot Dist(x_i, x_E) \tag{3}$$

where

$$Dist(x_i, x_E) = \exp\left(\frac{\left\| x_i - x_E \right\|^2}{Z_E}\right) \tag{4}$$

where Z_E is a normalization constant, and both ULD values and the Dist are normalized into the range of $[0,1]$.

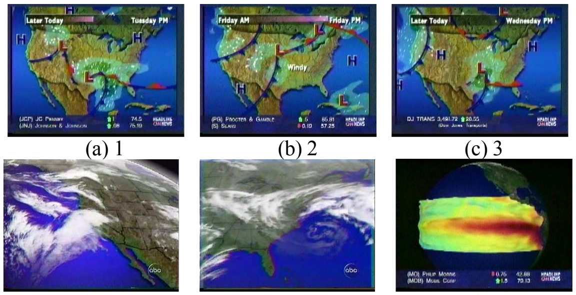
Based on the rank list generated above [6], a relevance learning scheme obtained from SVR is used to build models for general visual concepts.

III. Experimental Results

We demonstrate the performance of our algorithm using the NIST Video TRECVID 2003 corpus. The whole video dataset is divided into five parts: ConceptTraining, ConceptFusion1, ConceptFusion2, ConceptValidate, and ConceptTesting [12]. To show the performance of our Autonomous Learning mechanism, we use a small dataset ConceptValidate as the training set, which is about 6 hours includes 13 video sequences with 4420 shots/key frames. We try the Autonomous Learning mechanism to train models for 20 concepts and test them in the dataset ConceptFusion1, which is also 6 hours, including 13 news videos with 5,037 shots.

To show the performance of GMIL-ULD algorithm more clearly, we provide two algorithms as baselines for comparison. The first is described in [10], which is a supervised algorithm based on the Support Vector Machine. The second is an SVR-based confidence ranking, which are obtained from an ASR unimodal model described in [9]. We use the non-interpolated average precision over 1000 retrieved shots as a measure of the retrieval effectiveness, which is defined by NIST [12].

For the concept "Weather_News", there are 1696 Quasi-Positive bags based on the ASR unimodal analysis [9]. Using 576-bin Compressed-Domain Slice features [11], the GMIL-ULD algorithm provides relevance scores for each key frame. The top three and the lowest three pictures for "Weather_News" in the rank list are shown in Figure 3. We can see that the most informative visual model for this concept is close to (a), (b), and (c), whose influences are strongest among all the training data. While (f) is not so frequently shown for this concept, its influence to the model learning is weakened through the GMIL-ULD algorithm. Based on the obtained relevance score rank list, SVR is used to learn a regression model for "Weather_News".



(a) 1 (b) 2 (c) 3

(d) 26 (e) 46 (f) 80

Figure 3. Part of the Training Data for "Weather_News" with relevance score ranks based on GMIL-ULD (Note: The number below the picture shows the rank based on the relevance score.)

We get an average precision of 0.7446 for "Weather_News" and 0.092 for "Airplane" by using Autonomous Learning mechanism. We also trained and tested the two baseline algorithms using the same dataset. The results show that for "Weather_News", the average precision for SVM based supervised algorithm [10] is 0.4743, and for SVR based audio confidence score rank list [9], the average precision is 0.5265. For "Airplane", the average precisions for these two baseline algorithms are 0.0173 and 0.0114 respectively. Figure 4 and Figure 5 show the Precision-Recall curves. Fig. 6 compares the performances of SVM based supervised learning, ASR analysis based SVR, and Autonomous Learning algorithms by using the average precision for the concepts "Weather_News", and "Airplane". In these two examples, our algorithm outperforms both the SVM-based supervised learning algorithm and the SVR-based ASR analysis. Table 1 illustrates the average precisions for the 20 visual models.

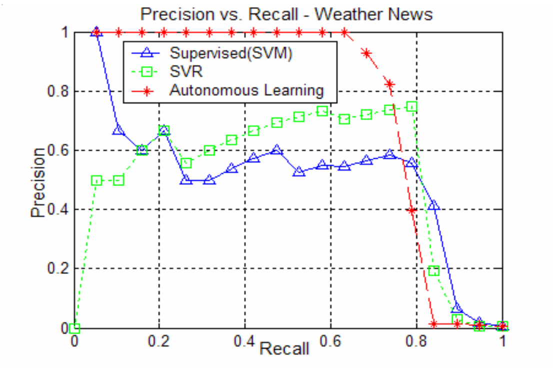


Figure 4. Performance comparison for "Weather_News"

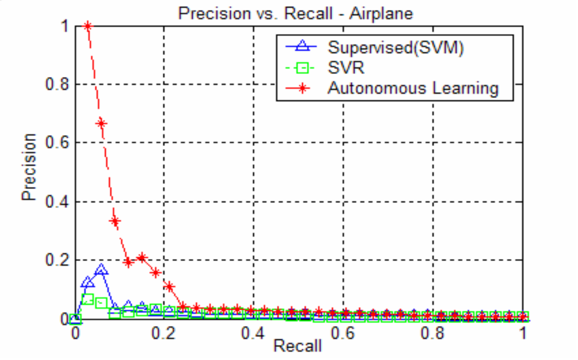


Figure 5. Performance comparison for "Airplane"

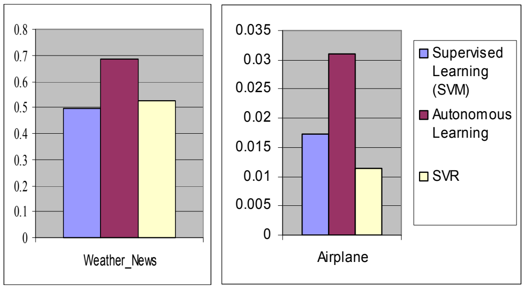


Figure 6. Average precision comparison

IV. Conclusions

We presented an Autonomous Learning mechanism and used it to build generic visual concept models from cross-modality information. Not only just reducing the number of examples needed for labeling, as active learning and transductive learning do, this whole process does not need any labeling work. Based on cross-modality information, Generalized Multiple Instance Learning and Uncertain Labeling Density are proposed for "Uncertainty Pruning". It is used to find the most informative example for the concept we are interested in. The relevance scores are calculated by considering the distance from this most informative example as well as the ULD value. This Autonomous Learning mechanism is tested for several concepts. In some preliminary experiments, our algorithm gives better or comparable performance to two baseline supervised algorithms -- an SVM based supervised learning algorithm and an audio confidence score rank list based SVR algorithm, Ongoing works include trying to trace the shot to find more suitable candidates for learning the recognition model and large-scale video concept modeling from broadcasting TV videos.

References

[1] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," Proc. of the ICML, New Brunswick, NJ, July 1994.

[2] T. Joachims, "Transductive inference for text classification using support vector machines" Proc. of ICML, pp. 200-209, 1999.

[3] Y. Rui, T. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," IEEE Transactions on Circuits and Systems for Video Technology, 8/5, pp. 644-655, 1998.

[4] X. Song and C.-Y. Lin and M.-T. Sun, "Cross-modality automatic face model training from large video databases ," Proc. of FPIV'04, Washington DC, June 28, 2004.

[5] X. Song, C.-Y. Lin and M.-T. Sun, "Autonomous Visual Model Building Based on Image Crawling through Internet Search Engines," Proc. of MIR 2004, New York, NY, October 2004.

[6] O. Maron, "Learning from ambiguity," PhD dissertation, Department of Electrical Engineering and Computer Science, MIT, Jun. 1998.

[7] T. G. Dietterich, R. H. Lathrop and T. Lozano-P_erez, "Solving the multiple instance problem with axis-parallel rectangles," Artificial Intelligence Journal, 89, pp. 31-71, 1997.

[8] R. A. Amar, D. R. Dooly, S. A. Goldman, and Q. Zhang, "Multiple-instance learning of real-valued data," Proc. of ICML, Williamstown, MA, pp. 3-10, 2001.

[9] C.-Y. Lin, M. R. Naphade, A. Natsev, C. Neti, J. R. Smith, B. Tseng, H. J. Nock, W. Adams, "User-trainable video annotation using multimodal cues," SIGIR, pp. 403-404, 2004.

[10] C.-Y. Lin, B. L. Tseng, M. Naphade, A. Natsev and J. R. Smith, "VideoAL: A Novel End-to-End MPEG-7 Automatic Labeling System," Proc. of ICIP, Barcelona, Sep. 2003.

[11] C.-Y. Lin, O. Verscheure and L. Amini, "Semantic Routing and Filtering for Large-Scale Video Streams Monitoring", submitted to Intl. Conf. on Multimedia and Expo (ICME), July 2005.

[12] A. Amir, *et. al.*, "IBM Research TRECVID-2003 Video Retrieval System," NIST TREC-2003 Video Retrieval Evaluation Conference, Gaithersburg, MD, Nov. 2003.

Table 1 Average Precision of 20 visual concepts

Concept	CF1	Concept	CF1
Weather_News	0.7446	Outdoor	0.0934
Human	0.6169	Airplane	0.092
Studio_Setting	0.4007	Sky	0.0829
Basketball	0.3186	Food	0.0789
Face	0.2342	Greenery	0.0498
Indoors	0.2001	Sport_Event	0.0446
Male_Face	0.1363	Water_Body	0.0354
People	0.1262	Building	0.0214
Graphics	0.1035	Road	0.0241
Crowd	0.0996	Hockey	0.0176