Multiple Instance Learning for Labeling Faces in Broadcasting News Video

Jun Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
juny@cs.cmu.edu

Rong Yan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
yanrong@cs.cmu.edu

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alex+@cs.cmu.edu

ABSTRACT

Labeling faces in news video with their names is an interesting research problem which was previously solved using supervised methods that demand significant user efforts on labeling training data. In this paper, we investigate a more challenging setting of the problem where there is no complete information on data labels. Specifically, by exploiting the uniqueness of a face's name, we formulate the problem as a special multi-instance learning (MIL) problem, namely exclusive MIL or eMIL problem, so that it can be tackled by a model trained with *partial* labeling information as the anonymity judgment of faces, which requires less user effort to collect. We propose two discriminative probabilistic learning methods named Exclusive Density (ED) and Iterative ED for eMIL problems. Experiments on the face labeling problem shows that the performance of the proposed approaches are superior to the traditional MIL algorithms and close to the performance achieved by supervised methods trained with complete data labels.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*

General Terms

Algorithm, Performance, Experimentations

Keywords

Face Labeling, News Video, Machine Learning, Multiple Instance Learning

1. INTRODUCTION

Labeling faces appearing in video with their names is of great significance to the better indexing and retrieval of broadcasting news video archive, which records the activities of a large number of important people. In previous works [2, 7, 11, 21, 20], face labeling is achieved by associating faces detected from video frames with people names extracted from video transcript (typically the closed-caption text) in the vicinity of the face. Specifically, in [21] we formulate this problem as a classification problem as to determine whether the association between a face and a candidate name is correct or incorrect. In this approach, each face-name association is treated as an data instance with a binary "correct/incorrect" label, and a classifier is trained from labeled data using a supervised method to predict the probability that a name is correctly associated with a face.

The supervised approach requires sufficient labeled training data in order to achieve high accuracy. This demands labeling the correct and incorrect names for a large number of sample faces, which is labor-intensive and time-consuming due to the huge volume of the video data and the lack of effective annotation tools. In this paper, we focus on a more challenging setting of the problem, where there is *no complete* knowledge on data labels available to the learning methods. Instead, we exploit the *partial* label information in the form of constraints on data labels which can be derived from the distinct properties of the problem. Specifically, since each face may have only one name, among all the face-name associations between a face and its candidate names, only one can be correct while all the others must be incorrect. It is also possible that none of the associations is correct, which means the face's name is not among the candidate names (i.e., anonymous face). From the machine learning perspective, this corresponds to a setting where unlabeled instances are assigned into a number of groups, and within each group, at most one instance is positive while all the others are negative.

Multiple Instance Learning (MIL) is a class of learning algorithms for handling problems with only partial label information expressed as the labels on *bags* of instances. In the MIL setting, unlabeled instances are grouped into a set of bags, and each bag is assigned a binary label which has a logical-or relationship with the instance labels. That is, if a bag is labeled positive, at least one instance in it is positive, and if a bag is labeled negative, all the instances in it are negative. Given the analogy between them, we can formulate face labeling as a MIL problem if (1) we treat each face-name association as an instance, and group the associations between a specific face and all its candidate names as a bag of instances; (2) we assign a label on each bag. As we will find out, here labeling a bag is equivalent to judging whether a specific face is anonymous or not, which requires much less user efforts than labeling all the instances

MM'05, November 6–11, 2005, Singapore.

in the bag. Therefore, it is appealing to tackle the face labeling problem using MIL methods especially when minimum manual effort is a preference.

Furthermore, since a face's name is unique, there is *exactly one* positive instance in each positive bag, denoted as *exclusive constraint*. This contrasts to the less informative constraint in the general MIL setting, where there is *at least one* positive instance in each positive bag. Therefore, face labeling belongs to a special type of MIL problem, named as *exclusive MIL* or eMIL problem. Consequently, the traditional MIL methods are not appropriate solutions to this eMIL problem since they do not take into account the exclusive constraints. In view of this, we propose two discriminative learning algorithms, Exclusive Density (ED) and Iterative ED, to solve eMIL problems and particularly our face labeling problem.

In this paper, we first review the related work in the area of face detection/identification in news video and multiple instance learning (Section 2). After that, we present an overview of the face labeling problem and discuss its formulation as an eMIL problem (Section 3). Two discriminative learning methods, ED and Iterative ED, are proposed for the eMIL problem (Section 4). Finally, we apply the proposed methods to the face labeling problem and compare their performance with both supervised learning methods and traditional MIL methods (Section 5). The experiments show that the performance of our approaches trained with only partial label information are not only superior to that of traditional MIL methods and but also close to that achieved by supervised methods trained with complete data labels.

2. RELATED WORK

In this section we review previous work in two areas, face detection/recognition in news video, which is the target application of this paper, as well as multi-instance learning, which inspires the formulation and solution to our problem.

2.1 Faces in News Video

As a critical step towards semantic video retrieval, video annotation with various semantic concepts has been an active research area in recent years. A number of methods [4, 18, 8, 17] were proposed to address a wide range of concepts such as those related to people (face, anchor), acoustic (speech, music, pause), object (buildings, graphics), location (outdoors, city, studio), genre (weather, financial, sports), and so on. Among them, the detection and recognition of faces is of great significance especially for broadcasting news video, which contains a large number of people as well as their activities. For example, Schneiderman et al. [12] has proposed a face detector that detects the faces appearing on video frames; Chen et al. [5] discussed robust face recognition in news video; Yang et al. [20] addressed the problem of finding named people in news video by combining clues from closed-captions, face similarity, and other semantic concepts. Recently there has been interest on associating faces with names in the news video [11, 21, 7], which is the main focus of this paper.

Name-It [11] is the first proposal on associating names with faces in news video, which is done by exploring the co-occurrence between faces in video frames and people names in video closed-captions. The underlying idea is that the (similar) faces that frequently co-occur with a certain name are likely to match the name, and vice versa. Though theoretically sound, the robustness of this method can be affected by the unreliable face similarity estimation in the low-quality news video, and no serious performance evaluation has been reported on this work. Named Faces system [7] built a database of named faces by recognizing the people names overlaid on video frames (usually below the faces) using video Optical Character Recognition (OCR). This approach is subject to the quality of video OCR, and more seriously, it cannot handle all the faces whose names are not overlaid on the screen.

In our previous work [21], a supervised approach has been adopted to identify correct associations between faces and names in news video based on multi-modal features. It builds a model from labeled data to predict the probability that a name in the closed-captions matches a face on the video frame, from which we are able to predict the name of each face. Although experimentally effective, this approach requires sufficient training data as a large number of sample faces manually labeled with names to achieve high accuracy. In this paper, we will investigate an alterative approach to the same problem which requires much less manual effort in collecting labeled data.

2.2 Multiple Instance Learning

Multiple instance learning (MIL) is proposed for problems with incomplete knowledge on data labels. Instead of receiving labeled instances as a conventional supervised method does, a MIL method receives a set of labeled *bags*, where each bag contains a number of unlabeled instances. If a bag is labeled positive, *at least one* of its instances must be positive, and if the bag is labeled negative, all of its instances must be negative. Thus, a bag label can be regarded as a constraint in the form of logical-or relationship with the labels of the instances in the bag. The goal of MIL is to derive a hypothesis from the bag labels to classify unseen instances and/or bags.

MIL was first introduced by Dietterich et al. [6] to solve the drug activity prediction problem, for which they proposed a class of methods for learning axis-parallel rectangles (APRs) as the target hypothesis. Besides, Diverse Density (DD) is a widely used method for MIL problems proposed by Maron et al. [9], and EM-DD was proposed by Zhang et al. [22] as an iterative variant of DD. Supervised learning algorithms have been adapted to MIL as well, such as the two SVM variants for MIL proposed by Andrews et al. [1], and the k-Nearest Neighbor method for MIL proposed by Wang et al. [16].

Besides the drug activity prediction problem [6], MIL has been used in a variety of applications. For example, content-based image retrieval and classification has been a popular application of MIL [23, 10, 19], where each image is modeled as a bag of regions and MIL methods are used to find the regions containing the target object. Similarly, MIL has been applied to text categorization by modeling a document as a bag of paragraphs [1]. Other applications include person recognition from images [15] and stock market analysis [9]. An interesting work closely related to this paper was done by Song et al. [14, 15], who used extended multi-instance learning methods to find faces or other visual objects from images returned by a search engine [14] or from video snippets returned by a news video retrieval system [15]. Here, the role of MIL is to distinguish the target face from the other faces that appear in the same image or video segment.

Figure 1: An illustration of face labeling in one story of the news video. Faces detected in monologue-speech shots are associated with people names found in the closed-captions.

3. FACE LABELING AS A MULTIPLE INSTANCE LEARNING PROBLEM

The goal of face labeling in news video is to label important faces appearing in the video with their names. Since news video usually comes with closed-caption text, which contains the names of the people involved in the news, the problem of face labeling boils down to finding correct associations between faces appearing in the video frames and the names found in the closed-captions. Moreover, a news video is temporally partitioned into a series of news stories, where each story consists of a sequence of video shots on the same event. Since each news story is an independent and semantically coherent unit, we assume that the name of a face, if mentioned in closed-captions, will only appear within the boundary of the same story that contains the face. Therefore, labeling a face in a story is to choose the most likely name from a set of candidate names in the closed-captions of that story.

Figure 1 illustrates the face labeling process in a news story. Faces are detected from video frames using a face detector [12], and meanwhile people names are extracted from closed-captions using a named-entity detector [3]. However, news video usually contains a large number of faces, some of which are anonymous and of little interest to users. For simplicity, we only label the faces of the people who are giving a monologue-style speech in news video. This not only reduces the number of faces, but also ensures that the remaining faces are worthwhile to label, because usually only the important people are given the chance to talk individually in the broadcasting news. The people in monologue speech can be identified either manually or automatically using the approach described in [13].

With the faces and names extracted, face labeling can be formulated as a machine learning problem as to estimating the probability that a given face is correctly associated with a candidate name, expressed as:

$$P(Y = 1|F, N, h) \tag{1}$$

where F and N represents face and name respectively, and $Y \in \{0, 1\}$ is a binary label indicating whether the face is associated with the name. Here h is the hypothesis which consists of a set of model parameters to be determined. In practice, each pair of F and N is described by a feature vector denoted as X, and therefore the probability of a name being associated with a face can be rewritten as:

$$P(Y = 1|F, N, h) = P(Y = 1|X, h) \tag{2}$$

Suppose f_i is a face to label, $\{n_{i1}, ..., n_{im_i}\}$ is the set of candidate names for f_i, and x_{ij} is the feature vector describing the association between f_i and n_{ij}. (Note that the set of candidate names is different for different faces, because for each face we only consider the names in the same story as the face.) If the hypothesis h is known, we can compute $P(Y = 1|x_{ij}, h)$ as the probability of n_{ij} being associated with f_i. Therefore, we can predict the name of the face f_i to be the one with the highest probability of being associated with f_i, expressed as:

$$n_i^* = arg \max_{n_{ij}} \log P(Y = 1|x_{ij}, h) \tag{3}$$

The feature vector x_{ij} for a face-name pair covers a variety of features derived from multiple video modalities, such as speaker identification, overlaid text, temporal video structure, and so on. Instead of presenting the details on how each feature is computed, we summarize all the features in Table 1. As we can see, each feature provides a clue as to how strongly a candidate name is associated with a face from a specific perspective. For example, the temporal distance between the position where the face appears and the position where the name is mentioned is informative since usually short distance indicates a likely match; the gender of the face derived from his/her voice and the gender of the name shows whether their gender matches; the type of face (as anchor, reporter, or news subject) helps match with the type of name to get ride of unlikely candidate names. Interested users may refer to [21] for a detailed description on the extraction of these features.

At the core of the above formulation is the learning of h. A natural solution is to apply a supervised learning model and estimate h from labeled training data in the form of $\{x_{ij}, y_{ij}\}$, where $y_{ij} \in \{0, 1\}$ is the manually assigned label indicating whether name n_{ij} is associated with face f_i or not. This supervised approach is exactly what we adopted

Table 1: A feature set describing the association between a *face* and a *name*

Modality	Feature	Description
Face Type	*type-match*	whether the predicted type of *shot* as anchor, reporter, or news-subject matches with the type of *name* derived by text analysis
Temporal Relationship	*face-name-dist*	the temporal distance (seconds) between *face* and the nearest occurrence of *name*
	name-rank	the rank of *name* among the all the names ranked by their distance to *face*
	face-name-order	whether *name* is mentioned before, within, or after *face*
Overlaid Text	*vocr-similarity*	the similarity between *name* and the video OCR output of the shot containing *face*
	vocr-present	whether there is overlaid text on the shot containing *face*
Speaker Identity	*utter-name*	whether *name* is uttered by the person of that *face*
	gender-match	whether the gender of *name* matches the gender of the person of *face*

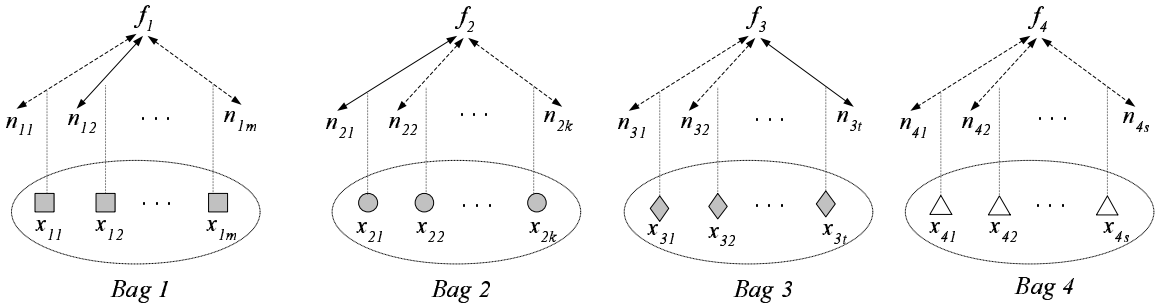


Figure 2: The formulation of face labeling as a multiple instance learning problem

in our previous work [21], which achieved reasonable performance in practice. To collect training data for this approach, however, we need to manually label the correct/incorrect names for a large number of sample faces. This means going through a portion of the video in a shot-by-shot manner to judge whether each name found in closed-captions matches each of the faces on the video frame, which is a tedious process. Therefore, it makes sense to explore a more challenging setting where there is no complete information on the data labels. Particularly, we seek for a formulation of face labeling as a special multiple instance learning (MIL) problem.

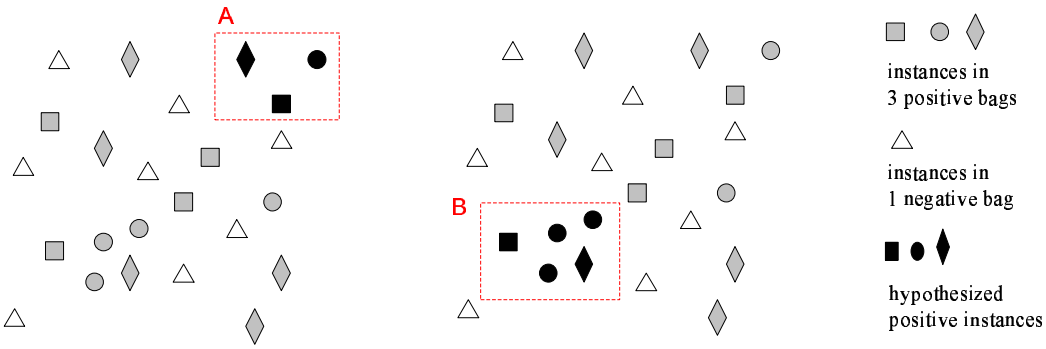
As shown in Figure 2, each face forms several *possible* associations with a set of candidate names, and each association maps to an data instance x_{ij}, whose label y_{ij} indicates whether the association is correct (positive) or incorrect (negative). Therefore, each face f_i corresponds to a set of instances as $\{x_{i1}, ..., x_{im_i}\}$, which can be naturally grouped into a *bag* of instances, shown by icons of the same shape. Each face belongs to either of the following two types: (1) a "non-anonymous" face, such as f_1, f_2 and f_3, for which one of its candidate names is the correct name, indicated by the solid-line link between them in Figure 2; (2) a "anonymous" face, such as f_4, for which none of its candidate names is correct. In the supervised setting, we know what the correct (and incorrect) names for each face are, and thus all the instance labels are available. Here, suppose we only know whether a face is anonymous or not, the information known as *bag labels*. That is, if a face is non-anonymous, there is *exactly one* positive instance in the corresponding bag (but we do not know which one), and in this case we say the bag is positive; if a face is anonymous, all instances in the bag are negative, and in this case we say the bag is negative. In Figure 2, instances from positive bags are shown as gray icons, while those from negative bags are shown as white icons. So the question is: can we drive the instance labels from the bag labels?

We formally define a bag x_i and its bag label y_i as:

$$\begin{aligned} x_i &= \{x_{i1}, ..., x_{im_i}\} \\ y_i &= y_{i1} \vee ... \vee y_{im_i} \end{aligned} \tag{4}$$

where x_{ij} is an instance corresponding to the association between face f_i and name n_{ij}, and y_{ij} is its (unknown) instance label. m_i is the number of instances in the bag, which is equal to the number of candidate names for face f_i. Our goal is to learn a hypothesis h from the given bag labels $\{y_i\}$ that can predict the unknown instance labels $\{y_{ij}\}$. This exactly fits into the setting of multiple instance learning (MIL) described in Section 2.2. Therefore, a variety of existing MIL approaches [6, 22, 9] can be readily applied to find such a hypothesis and solve the face labeling problem.

Using MIL methods for face labeling is appealing only if the bag labels as the anonymity of sample faces are less expensive to collect than the instance labels as the correct/incorrect names for the sample faces. This is exactly the case in our problem, since there exist many clues to easily determine the anonymity of a face in monologue speech. In news video, the name of a monologue face usually appears either in the closed-captions or in the overlaid screen text, but very rarely in neither of the places. The faces with names as overlaid text can be easily distinguished using the video OCR techniques. Therefore, most (if not all) of the remaining faces have their names in the closed-captions, so they are "non-anonymous faces" and correspond to positive bags in our formulation. If we only focus on these faces, the bag labels are readily available without manual effort.



(a) hypothesis A valid for MIL and eMIL (b) hypothesis B valid for MIL but not eMIL

Figure 3: An example data set in 2-d feature space, where icons of the same shape represent instances from the same bag. Gray icons are instances from positive bags and white ones are instances from negative bags. Two hypotheses A and B are shown as rectangles, and the instances inside each rectangle are classified as (hypothesized) positive instances while those outside are classified as negative.

Note that not labeling the faces with overlaid names does not hurt the generality of our approach much, because their names can be easily labeled using video OCR techniques. Besides this clue, there are other clues that help determine the anonymity of a monologue face, e.g., faces appearing for a long duration or for many times are rarely anonymous, and so on. Therefore, obtaining bag labels in this problem demands significantly less human efforts and time than obtaining the complete instance labels.

A careful comparison reveals an important difference between face labeling and the standard MIL problem. In the face labeling problem, there is *exactly one* positive instance in each positive bag, since each face may have only one name. We call this "exclusive constraint", which is more informative than the logical-or constraint in the standard MIL setting, where there are *at least* one instance in a positive bag. Thus, we consider face labeling as a *special* MIL problem that contains extra information in the form of exclusive constraints. We call such MIL problems exclusive MIL or eMIL problems. Apparently, traditional MIL methods [6, 22, 9] are not appropriate solutions to an eMIL problem, since they overlook the additional information in exclusive constraints. (However, this is not the drawbacks of these MIL methods because they are for general MIL problems which do not necessarily satisfy exclusive constraints.) Actually, using MIL methods for an eMIL problem may pose mistakes in the search for optimal hypothesis. For example, Figure 3 shows a projection of the data instances from Figure 2 in a 2-D feature space, with two possible hypotheses A and B shown as rectangles. The instances inside each rectangle are classified as "genuinely" positive instances and those outside are classified as negative instances. According to the definition, both A and B are valid hypotheses for a MIL problem because each of them include at least one instance from each positive bag and meanwhile exclude all instances from negative bags. For an eMIL problem, however, B is only longer a valid hypothesis since it violates the exclusive constraints by classifying more than one instance in a positive bag as positive, whereas A is still valid since exactly one instance in each positive bag is classified positive. Thus, if applying MIL methods to this data-set, there is a high chance of concluding to the wrong hypothesis B, which may result in poor performance of the face labeling problem. In view of this, we discuss two discriminative learning methods, Exclusive Density (ED) and Iterative ED, which take advantage of the exclusive constraints in eMIL problems and therefore avoid mistakes as those discussed above.

4. LEARNING METHODS

We start this section by reviewing a widely-used general MIL approach *Diverse Density* (DD), and then propose *Exclusive Density*(ED) as a discriminative learning method for eMIL problems. Finally, an iterative variant of ED is presented.

4.1 Notations

The data D includes $\{x_1, ..., x_n\}$ as a set of n bags and $\{y_1, ..., y_n\}$ as their bag labels, where $y_i \in \{0, 1\}$. Each bag x_i consists of a set of instances as $x_i = \{x_{i1}, ..., x_{in_i}\}$, where x_{ij} denotes the j^{th} instance in bag x_i. Each instance is described by a feature vector, and x_{ijd} denotes the value of the d^{th} feature component of instance x_{ij}. The labels of instances in bag x_i are $\{y_{i1}, ..., y_{in_i}\}$. In the standard MIL setting, instance labels are bounded by the bag label through a logical-or relationship, i.e., $y_i = y_{i1} \vee ... \vee y_{in_i}$. In eMIL problems, there is a stronger constraint as $y_i = \sum_j y_{ij} \in \{0, 1\}$ due to the exclusive constraints. For simplicity, we denote the probability of bag label $P(Y = y_i|y_i, h)$ as $Pr(+|y_i, h)$ if $y_i = 1$, or as $P(-|x_i, h)$ if $y_i = 0$. Similarly, we denote the probability of instance label $P(Y = y_{ij}|x_{ij}, h)$ as $Pr(+|x_{ij}, h)$ if $y_{ij} = 1$, or as $P(-|x_{ij}, h)$ if $y_{ij} = 0$. In our approach, a hypothesis $h = (\mu, \sigma)$ consists of two parts: the coordinates of a concept point as μ, and a scale vector σ defining the weights on different feature dimensions.

4.2 Diverse Density

Diverse Density (DD) [9] is a widely used MIL method. The underlying idea of DD is very intuitive: it tries to find a concept point in the feature space so that at least one instance from each positive bag is close to it, while all the instances from negative bags are far away from it. Thus, the diverse density as the "goodness" metric of a concept point is measured by how many positive bag has at least one instance close to it, and how far the instances

in negative bags are away from it. Formally, the diverse density of a hypothesis h is defined as the data likelihood $DD(h) = P(h|x_1^+, ..., x_n^+, x_1^-, ..., x_m^-)$ and the optimal hypothesis can be found by maximizing $DD(h)$ over the hypothesis space. Note that here we follow the original notation in [9], where x_i^+ and x_i^- denotes a positive or a negative bag, and x_{ij}^+ or x_{ij}^- denotes the j^{th} instance from a positive or a negative bag. With the assumption of uniform prior on the hypothesis space and the independence between bags given the hypothesis, this is equivalent to maximize $\prod_i P(x_i^+|h) \prod_i P(x_i^-|h)$. By applying the Bayes' rule, we have the following:

$$h_{DD} = arg\max_{h \in H} \prod_i P(h|x_i^+) \prod_i P(h|x_i^-) \tag{5}$$

where $P(h|x_i^+)$ and $P(h|x_i^-)$ are defined by $P(h|x_{ij}^+)$ and $P(h|x_{ij}^-)$ using noise-or model:

$$P(h|x_i^+) = 1 - \prod_j (1 - P(h|x_{ij}^+)) \tag{6}$$

$$P(h|x_i^-) = \prod_j (1 - P(h|x_{ij}^-)) \tag{7}$$

The causal probability $P(h|x_{ij})$ is estimated by Gaussian-like distribution $exp(-\sum_d (1/\sigma_d)^2 (x_{ijd} - \mu_d)^2)$, which is inversely related to the Euclidean distance between x_{ijd} and μ. The optimal h is found by gradient search using Eq(5) as the objective function.

4.3 Exclusive Density (ED)

Inspired by DD, we propose Exclusive Density (ED) for eMIL problems where exclusive constraints exist. Similar to the intuition of DD, ED tries to find a concept point in the feature space so that *exactly one* instance from each positive bag is close to it, while all the instances from negative bags are far away from it. Here, the "goodness" of a concept point is measured by how many positive bag has *exactly one* instance close to it, and how far the instances in negative bags are away from it. One can see that the key difference between DD and ED is on whether the optimal concept point should be close to "at least one" or "exactly one" positive instance in each positive bag. This echoes the difference between MIL and eMIL problem.

Formally, we define the exclusive density [1] of a hypothesis as the conditional likelihood of the bag labels given the data and the hypothesis, expressed as $ED(h) = L(h; D) = P(y_1, ..., y_n|x_1, ..., x_n, h)$ The optimal hypothesis h_{ED} can be found by maximizing this conditional likelihood, i.e., $arg\max_h L(h; D)$. Under the assumption of the independence between bag labels given the data and the hypothesis, this is equivalent to:

$$\begin{aligned} h_{ED} =& arg\max_{h \in H} \prod_i P(y_i|x_i, h) \\ =& arg\max_{h \in H} \prod_{\forall i: y_i = 1} P(+|x_i, h) \prod_{\forall i: y_i = 0} P(-|x_i, h) \end{aligned} \tag{8}$$

To represent Eq(8) as a function of h, the first step is to transform the probability on bag labels into probability on instance labels. This requires different transformations for positive and negative bags. With the exclusive constraints, a positive bag has only one positive instance and the rest are all negative instances. Therefore, the probability of a bag's label being positive depends on how likely one instance in the bag generates a positive label and meanwhile how likely the other instances generate negative labels. This leads to an intuitive definition as:

$$P(+|x_i, h) = \frac{1}{Z_i} \max_j \{P(+|x_{ij}, h) \prod_{k \neq j} P(-|x_{ik}, h)\} \tag{9}$$

where Z_i is a normalization factor. Since we do not know which is the only positive instance in the bag, the max() function intends to test all possible configurations and find the largest separation between any single instance and the other instances in the bag. However, since max() is not differentiable, no optimization methods can directly work on Eq(9). Therefore, we adopt the same noise-or model used in DD as a differentiable, soft-max function, which is expressed as $\max(x_1, .., x_n) \approx 1 - \prod_i (1 - x_i)$. Thus, we approximate Eq(9) by the following equations:

$$\begin{aligned} & P(+|x_i, h) \\ &= \frac{1}{Z_i} \{1 - \prod_j (1 - P(+|x_{ij}, h) \prod_{k \neq j} P(-|x_{ik}, h))\} \end{aligned} \tag{10}$$

Since all the instances in a negative bag must be negative, the probability that a bag generates a negative label is high if every instance in the bag has a high probability of being negative, which leads to:

$$P(-|x_i, h) = \frac{1}{Z_i} \prod_j P(-|x_{ij}, h) \tag{11}$$

The normalization factor Z_i in Eq(10) and Eq(11) is added to ensure well-defined probabilities, i.e., $\sum_{y_i} P(y_i|x_i, h) = 1$. Therefore, we have:

$$\begin{aligned} Z_i =& \{1 - \prod_j (1 - P(+|x_{ij}, h) \prod_{k \neq j} P(-|x_{ik}, h))\} \\ &+ \prod_j P(-|x_{ij}, h) \end{aligned}$$

According to Eq(10) and Eq(11), for a concept point to receive a high exclusive density $ED(h)$, it must be close to *exactly* one instance in each positive bag and meanwhile far away from the other, and also far away from the instances in negative bags.

To define $P(+|x_{ij}, h)$, we conceive each instance label y_{ij} as a random variable drawn from a Bernoulli distribution parameterized by the data point x_{ij} and the hypothesis $h = (\mu, \sigma)$. Specifically, we define

$$\begin{aligned} P(+|x_{ij}, h) &= f(\| x_{ij} - \mu \|_p) \\ P(-|x_{ij}, h) &= 1 - P(+|x_{ij}, h) \end{aligned} \tag{12}$$

where $\| x_{ij} - \mu \|_p$ is the p-norm distance between x_{ij} and the concept point μ, and $f(\cdot)$ is a function transforming the distance into a probability. Apparently, $f(\cdot) \in [0, 1]$ and it should be inversely related to the distance so that an instance closer to the concept point has a higher probability of generating a positive label. This formulation implies that, there is a Bernoulli distribution at each point of the feature space, and the label of an instance is drawn from the distribution at the corresponding point.

[1] We name our method as exclusive density due to its close connections with Diverse Density. It does not mean a density function.

Although f can be any function satisfying the above requirements, we define $f(x) = \exp(-x)$ and use the generalized L-2 distance metric, which leads to a Gaussian-like from:

$$P(+|x_{ij}, h) = \exp(-\sum_d \frac{(x_{ijd} - \mu_d)^2}{2\sigma_d^2}) \quad (13)$$

The final optimization function is given by plugging Eq(10) and Eq(11) into Eq(8). Since it is differentiable, we use gradient search method with multiple starting points to find h that maximizes $ED(h)$. Intuitively, the instances from every positive bag are good starting points, since some of them (actually, one in each bag) should be close to the true concept.

After h has been optimized, we can use it to predict the instance labels in the data. Under the exclusive constraints, only one instance in a positive bag can be positive, so it has to the one with the highest probability of generating a positive label:

$$y_{ij} = \begin{cases} 1 & \text{if } j = arg\max_k P(+|x_{ik}, h) \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

According to the definition of negative bags, all the instances in a negative bag are set to negative.

In practice, we made two modifications on the ED algorithm. First, since the normalization factors Z_i significantly increases the model complexity and the running time, we removes it from the optimization function of ED. We found experimentally this does not cause any non-trivial changes of the performance. The second modification is based on our observation that ED has a bias towards the instances that are likely to be negative more strongly than to the positive ones. One major reason is that the probability of a positive bag label as defined in Eq(9) is dominated by negative instances. As a result, the scale factors may become extremely large in order to fit the negative instances well, which sometimes causes numerical difficulties in the optimization process. To alleviate this problem, we modify the probability of a positive bag label in Eq(9) into $P(+|x_i, h) = \frac{1}{Z_i} \max_j P(+|x_{ij}, h) \prod_{k \neq j} P(-|x_{ik}, h)^{\gamma_i}$, where γ_i is introduced to balance off the contribution of positive and negative instances. Similar modification is made to Eq(10). Intuitively, in our experiments we set $\gamma_i = \frac{1}{n_i - 1}$ with n_i being the number of instance in x_i^+, so that the contribution from the two sides are made equal.

4.4 Iterative ED

The ED algorithm uses gradient search to optimize a complex objective function which leads to a very inefficient implementation. In this section, we present an efficient iterative method called *Iterative ED* which maximizes a simplified objective function that is related to ED. The idea was inspired by the notion that the difficulty of an eMIL problem mainly comes from the ambiguity of not knowing which instance is the only positive instance in a positive bag. In the ED algorithm, this ambiguity is modeled by the $max()$ function in Eq(9), which is then replaced by a very complicated noise-or model in Eq(10) as its differentiable approximation. The same problem exists in the standard MIL setting, for which Zhang et al. [22] have proposed a EM-style variant of DD called EM-DD to tackle the ambiguity by modeling the knowledge on which instance is positive in a bag using a set of missing variables.

Algorithm 1 Iterative ED: An iterative variant of ED

Input: a set of bags and the labels as $\{\langle X_1, y_1\rangle, ..., \langle X_n, y_n\rangle\}$
Output: a hypothesis $h(\mu, \sigma)$.

1: Choose an initial hypothesis $h^{(0)}$.
2: Randomly initialize the instance labels $\{y_{ij}^{(0)}\}$.
3: Choose ϵ as the minimum increase of the objective function for each iteration.
4: $Diff = MaxInteger$
5: $t = 0$
6: **while** $Diff > \epsilon$ **do**
7: **for** each positive bag x_i **do**
8: $y_{ij}^{(t+1)} = \begin{cases} 1 & \text{if } j = arg\max_k P(+|x_{ik}, h^{(t)}) \\ 0 & \text{otherwise} \end{cases}$
9: **end for**
10: $h^{(t+1)} = arg\max_{h \in H} CED(\{y_{ij}^{(t+1)}\}, h)$
11: $Diff = CED^{(t+1)} - CED^{(t)}$
12: $t = t + 1$
13: **end while**
14: Return $h^{(t)}$

Iterative ED uses an idea similar to EM-DD. Suppose we know which instance is the only positive instance in each positive bag under the current hypothesis $h^{(t)}$, and thus all the other instances are all negative. This means all the instance labels $\{y_{ij}^{(t)}\}$ are available. Therefore, we directly optimize our hypothesis by maximizing the data likelihood defined on instance labels instead of on bag labels. We call this simplified likelihood function as Complete-data Exclusive Density (CED), since it supposes the unknown instance labels $\{y_{ij}\}$ are available. It is expressed as:

$$CED^{(t)} = CED(\{y_{ij}^{(t)}\}, h^{(t)}) = L(h^{(t)}; D^{(t)})$$
$$= \prod_{\forall i,j: y_{ij}^{(t)}=1} P(+|x_{ij}, h^{(t)}) \prod_{\forall i,j: y_{ij}^{(t)}=0} P(-|x_{ij}, h^{(t)}) \quad (15)$$

By maximizing this function over the hypothesis space, we will get a new hypothesis, which in turn can be used to update instance labels. Iterative ED takes the advantage of the cyclic nature of this process. Starting from an initial hypothesis, it repeatedly performs the following two steps: In first step, it computes the conditional probability $P(+|x_{ij}, h^{(t)})$ of the instances in each positive bag, and set their labels $\{y_{ij}^{(t)}\}$ such that the one with the highest probability is labeled positive and the rest are labeled negative; In the second step, it updates the hypothesis $h^{(t+1)}$ by maximizing $CED^{(t)}$ using gradient search. Then we replace $h^{(t)}$ by $h^{(t+1)}$ and repeat the two steps until the algorithm converges. The pseudo-code of this process is shown in Algorithm 1. Its convergence is proved below.

PROPOSITION 1. *Iterative ED algorithm converges.*

PROOF. The convergence of the algorithm is proved by showing (1) the objective function $CED^{(t)}$ is monotonically increasing, and (2) it has an upper bound. (2) is obvious since $CED^{(t)}$ is defined as a product of probabilities in Eq(15) and thus $CED^{(t)} \leq 1$. To prove (1), note that $CED^{(t)}$ is completely defined by a hypothesis $h^{(t)}$ and the instance labels as $\{y_{ij}^{(t)}\}$. Hence, in each iteration, $CED^{(t)}$ may change only at Step 8 where $\{y_{ij}^{(t)}\}$ is updated or at Step 10 where $h^{(t)}$ is updated. In Step 10, the increase of $CED^{(t)}$

is guaranteed by the gradient ascent algorithm. In Step 8, $CED^{(t)}$ changes when the instance labels in at least one positive bag are different from the last step. Without loss of generality, we assume that only positive instance in bag x_i changes from x_{ij_1} in step t to x_{ij_2} in step $t+1$, and that is the only change happened in Step 8. This means $y_{ij_1}^{(t)} = 1$ and $y_{ij_2}^{(t)} = 0$ in step t, while $y_{ij_1}^{(t)} = 0$ and $y_{ij_2}^{(t)} = 1$ in step $t+1$. This change also implies $P(+|x_{ij_2}, h) > P(+|x_{ij_1}, h)$, based on which we derive:

$$\frac{CED^{(t+1)}}{CED^{(t)}} = \frac{P(+|x_{ij_2}, h)P(-|x_{ij_1}, h)}{P(+|x_{ij_1}, h)P(-|x_{ij_2}, h)} = \frac{P(+|X_{ij_2}, h)(1 - P(+|X_{ij_1}, h))}{P(+|x_{ij_1}, h)(1 - P(-|x_{ij_2}, h))} > 1$$

which shows $CED^{(t+1)} > CED^{(t)}$. Thus, the monotonic increase of $CED^{(t)}$ is proved. □

Although not a strict EM algorithm, Iterative ED works in an EM-style procedure that estimates the instance labels as missing variables and updates the hypothesis to maximize the likelihood in an interleaved manner. As pointed out by [22], such an iterative algorithm may help avoid being trapped in local minima since the algorithm makes sharp changes in the hypothesis when its guess of the instance labels changes. Further, Iterative ED is more efficient because it has a much simpler objective function than that of ED.

5. EXPERIMENTS

In this section, we first describe the test data set and the experiment set-up for the face labeling problem. Then we present the performance of the proposed learning methods and compare it with other methods.

5.1 Data Set

The test data of the face labeling problem are collected from the news video archive of ABC World News Tonight in 1998. Specifically, we use a collection of the news video in 20 days with 30 minutes per day. In the preprocessing stage, faces are detected from video frames using a face detector [12], and people names are extracted from closed-captions using a named-entity detector [3]. As mentioned in Section 3, we only label the faces in monologue-style speech, which can be automatically identified by a monologue detector [21]. Faces of anchors and reporters are further removed since they are not interesting to users. Finally, there are 476 monologue faces to label, among which 234 faces have their names found in the closed-captions (i.e., non-anonymous), and the other 242 faces are simply anonymous. In average there is 4.7 candidate names for every face to label. As discussed in Section 3, we treat each face-name association as an instance, and group the associations (instances) related to the same face into a bag. This results in 476 bags (234 positive and 242 negative) consisting of totally 2236 instances. Each instance as a face-name association is described by a 8-dimensional feature vector described in Table 1. The work described in this paper has been also incorporated into a news video browsing and retrieval system. As shown in Figure 4, the interface of the system displays bounding boxes around the detected faces in each video shot. When the cursor is moved over a face, it shows two most likely names of that face predicted using the methods proposed here.

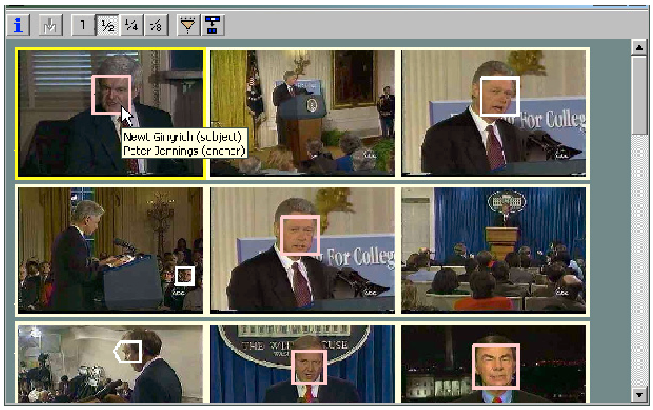


Figure 4: Interface showing the predicted names of a face

5.2 Experiment Set-up

We conduct the experiments in two settings. In the first setting, `PosOnly`, only the 1284 instances from 234 the positive bags are used, which means we try to label only the non-anonymous faces. For comparison purpose, we evaluate the performance of 6 algorithms: `ED` and `IterED` are the exclusive density and iterative ED method proposed for eMIL problem; `DD` [9] and `EM-ED` [22] are widely used traditional MIL algorithms; `LDA` and `SVM` are two supervised learning algorithms, namely linear discriminant analysis and support vector machine (with RBF kernel). The first 4 algorithms work only with the bag labels (which are all positive in this setting) and thus it is legitimate to use the same data-set for both training and testing; LDA and SVM work in supervised setting with all instance labels available, and 10-fold cross-validation is used to evaluate their performance. The two supervised methods are included to give an upper bound of the performance, so that we can see how close (to this upper bound) our methods can achieve without knowing the instance labels. We label a face using the name which receives the highest probability of being associating with the face. The performance is evaluated by the accuracy of predicted names, which is the ratio of the correctly labeled faces against all the faces. Note that this is different from the classification accuracy of instance labels, since a bag of correctly labeled instances (names) only translates into one correctly labeled face. In the second setting, `AllData`, we use all the 2236 instances from both positive and negative bags. This includes both the anonymous faces and the non-anonymous faces. In this case, since the instances in negative bags are all known to be negative, we cannot use the same data-set for testing and training. Therefore, we use 10-fold cross-validation to evaluate the performance of all the 6 algorithms.

We ran `ED`, `IterED`, `EM`, and `EM-DD` all with 50 starting points, where each starting point is the feature vector of an instance randomly chosen from positive bags. This brings up the issue of combining the hypotheses resulted from the multiple runs. We use two strategies, `max`, which uses the single hypothesis that achieves the maximum value of the data likelihood function and discards the other hypotheses, and `avg`, which keeps all the hypotheses and computes the probability of an instance's label as the average of its probabilities predicted under all the hypotheses.

Table 2: Comparison on accuracies of face labels

Algorithm	Hypotheses Selection	Accuracy of face labels	
		PosOnly	AllData
ED	avg	**0.590**	**0.597**
	max	**0.590**	0.593
IterED	avg	0.543	0.573
	max	0.577	0.592
DD	avg	0.491	0.548
	max	0.449	0.568
EM-DD	avg	0.470	0.440
	max	0.478	0.502
LDA (supervised)		0.606	0.621
SVM (supervised)		0.616	0.631

Table 3: Comparison of average running time

Algorithm	Average running time (sec)
ED	222.4
IterED	127.1
DD	105.0
EMDD	4.7

5.3 Results

Table 2 summarizes the accuracies of face labels predicted by the 6 algorithms under the `PosOnly` and the `AllData` setting. Several interesting observations can be made. First, in the `PosOnly` setting, ED and IterED outperforms DD and EM-DD by a large margin. This indicates the importance of exploiting the exclusive constraints in the eMIL problem. Second, in the `AllData` setting, due to the availability of additional instances with known negative labels, the performance of almost all the algorithms improves. However, the improvement of ED/IterED is not very significant, and their advantage over DD/EM-DD shrinks compared with that in the `PosOnly` setting. This implies that our ED/IterED algorithm does a better job than DD/EM-DD in terms of making use of the unlabeled data so that the additional (labeled) data do not help much. Third, the performance of ED, which is trained with only the bag labels, approaches the performance of LDA and SVM, which are trained with fully labeled instances in supervised setting. This is a significant observation which means the proposed methods can solve the face labeling problem almost as well as the supervised methods while requiring less user efforts in collecting training data labels. Fourth, despite the efficiency issue, the two iterative variants, i.e., IterDD and EM-DD, work slightly worse than their original versions. Fifth, there is no strong evidence preferring `avg` or `max`, since neither of them is consistently better than the other.

We also compare the efficiency of the proposed algorithms with the traditional MIL methods. Table 3 summarizes the average time needed for running each algorithm in the `AllData` setting using a single starting point. As expected, the ED algorithm is about two times slower than DD, because its optimization function is more complicated than the latter one, although both of them use gradient descent as the optimization method. A bit surprisingly, between the iterative variants of these two algorithms, IterED does not enhance the efficiency (from ED) as much as EM-DD does. A possible explanation is that, even after removing the $softmax$ function in the optimization function, IterED's objective function still involves the probabilities of all the instances and therefore is rather complex.

6. CONCLUSION

We have investigated the problem of face labeling with only partial information on training data labels. Specifically, we have presented a formulation of the problem as an exclusive MIL or eMIL problem, and proposed two discriminative learning methods to address such problems. The effectiveness of the proposed methods has been demonstrated by the experiments on the face labeling problem, where their performance is superior to the traditional MIL algorithms and close to that achieved by supervised methods trained with complete data labels.

Although the proposed methods are applied only to the face labeling data-set, they are applicable to any other problems that can be formulated as eMIL problems. Such problems can be frequently observed in many video/image applications, in which the exclusive constraints are available because of the *unique identity* of real-world objects. We give some examples below:

- In the surveillance video of a hospital, there is often a need to recognize the identity of the patients captured in the video to monitor their behaviors. Since each person may correspond to only one patient-ID, this problem can be formulated as an eMIL problem if we treat the associations between an observed person and the possible patient-IDs as a bag of instances.
- To automatically generate a meeting minute from the video recording of a meeting, one needs to identify the speaker behind each voice. Since one voice has only one speaker-ID, this is also an eMIL problem if the association between a voice and the possible speaker-IDs are considered as a bag of instances.
- A user of an image retrieval system may want to find a unique object, say, "Statue of Liberty", from a set of images, where each image is partitioned into several regions. If we treat regions of an image as a bag of instances, this is also an eMIL problem since "Statute of Liberty" can appear in only one of the regions.

Besides exploring the above problems, we can also improve the proposed methods which are very limited in terms of representation power since they are unable to model complicated class boundaries. Therefore, we plan to adapt more sophisticated supervised learning methods such as SVM to eMIL problems.

7. ACKNOWLEDGEMENTS

This work is supported in part by the Advanced Research and Development Activity (ARDA) under contract numbers NBCHC040037 and H98230-04-C-0406.

8. REFERENCES

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, 2003.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in news. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 848–854. IEEE Computer Society, 2004.

[3] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder, 1997.

[4] S. F. Chang, R. Manmatha, and T. S. Chua. Combining text and audio-visual features in video indexing. In *IEEE ICASSP 2005*, 2005.

[5] M. Chen and A. Hauptmann. Toward robust face recognition from multiple views. In *Proc. of Int'l Conference on Multimedia and Expo*, 2004.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[7] R. Houghton. Named faces: Putting names to faces. *IEEE Intelligent Systems*, 14(5):45–50, 1999.

[8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th Int'l ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 119–126, 2003.

[9] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, volume 10, pages 570–576. The MIT Press, 1998.

[10] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proc. 15th Int'l Conf. on Machine Learning*, pages 341–349. Morgan Kaufmann, 1998.

[11] S. Satoh and T. Kanade. Name-it: Association of face and name in video. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 368–373. IEEE Computer Society, 1997.

[12] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *Int. J. Comput. Vision*, 56(3):151–177, 2002.

[13] C. Snoek, M. Worring, and A. Hauptmann. Detection of TV news monologues by style analysis. In *Proc. of the IEEE Int'l Conference on Multimedia & Expo*, June 2004.

[14] X. Song, C.-Y. Lin, and M.-T. Sun. Autonomous visual model building based on image crawling through internet search engines. In *Int'l Workshop on Multimedia Information Retrieval*, pages 315–322. ACM Press, 2004.

[15] M.-T. S. Song Xiaodan, Ching-Yung Lin. Cross-modality automatic face model training from large video databases. In *Workshop on Face Processing in Video*, 2004.

[16] J. Wang and J.-D. Zucker. Solving the multiple instance problem: A lazy learning approach. In *Proc. 17th Int'l Conf. on Machine Learning*, pages 1119–1125. Morgan Kaufmann, 2000.

[17] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proc. of the 12th annual ACM Int'l Conf. on Multimedia*, pages 572–579, 2004.

[18] R. Yan and M. R. Naphade. Semi-supervised cross feature learning for semantic concept detection in video. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2005.

[19] C. Yang and T. Lozano-Perez. Image database retrieval with multiple-instance learning techniques. In *Proc. of Internatinal Conf. on Data Engineering*, pages 233–243, 2000.

[20] J. Yang, M. Chen, and A. G. Hauptmann. Finding person X: Correlating names with visual appearances. In *Proc. of 3rd Int'l Conf. on Image and Video Retrieval*, pages 270–278, 2004.

[21] J. Yang and A. G. Hauptmann. Naming every individual in news video monologues. In *Proc. of the 12th annual ACM Int'l Conf. on Multimedia*, pages 580–587. ACM Press, 2004.

[22] Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems*, pages 1073–1080. The MIT Press, 2001.

[23] Q. Zhang, S. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple-instance learning. In *Proc. 19th Int'l Conf. on Machine Learning*, pages 682–689, 2002.