Learning the Semantics of Images by Using Unlabeled Samples

Jianping Fan, Hangzai Luo, Yuli Gao
Department of Computer Science, UNC-Charlotte, NC28223, USA

Abstract

In this paper, we have proposed a novel framework to achieve more effective classifier training by using unlabeled samples. By integrating concept hierarchy for semantic image concept organization, a hierarchical mixture model is proposed to enable multi-level image concept modeling and hierarchical classifier training. To effectively learn the base-level classifiers for the atomic image concepts at the first level of the concept hierarchy, we have proposed a novel adaptive EM algorithm to achieve more effective classifier training with higher prediction accuracy. To effectively learn the classifiers for the higher-level semantic image concepts, we have also proposed a novel technique for classifier combining by using hierarchical mixture model. The experimental results on two large-scale image databases are also provided.

1. Introduction

As high-resolution digital cameras become more affordable and widespread, high-quality digital images have exploded on the Internet. With the exponential growth on high-quality digital images, the need for automatic image annotation is becoming increasingly important in order to support semantic image retrieval via keywords [1].

Semantic image classification is a promising approach to enable automatic image annotation and becomes increasingly attractive [2-7]. However, one major difficulty for most existing techniques is that a large number of labeled samples are required to learn the concept models accurately. Unfortunately, labeling a large number of training images is very expensive and time-consuming.

Given this costly labeling problem, it is very attractive to design techniques that can take advantage of unlabeled samples for semi-supervised classifier training [8-10]. One basic assumption for most existing semi-supervised techniques is that each unlabeled sample originates from one of the known image context classes that can be effectively learned from the available labeled samples, where each known image context class is represented by using one or multiple mixture components [4-5]. When only a limited number of labeled samples are available for classifier training, this basic assumption is not satisfied because of concept uncertainty (i.e. presence of new concept, outliers, and unknown image context classes) [9]. Considering that many image context classes and new concept have not even occurred in a limited number of labeled samples, using the outlying unlabeled samples will corrupt the density estimation and lead to worse performance rather than improvement when the underlying model structure is incorrect.

Through a density-based approach, finite mixture model has provided a natural way to deal with the problem of semi-supervised classifier training by using both the labeled samples and the unlabeled samples [16]. One major shortcoming of finite mixture model is that negative samples are not incorporated for classifier training. Another shortcoming for most existing techniques is that they ignore the hierarchical relationships among different semantic image concepts by independently learning a set of flat classifiers. Some pioneer works have shown that the accuracy of text classifiers may be significantly improved by taking advantage of concept hierarchy for classifier training [17-18].

Based on these observations, we have developed a new framework to achieve more effective hierarchical classifier training. This paper is organized as follows: Section 2 presents a novel framework for hierarchical image concept modeling and organization; Section 3 proposes a novel algorithm for classifier training; Section 4 shows our work on algorithm evaluation; We conclude in Section 5.

2. Hierarchical Concept Modeling

To achieve more effective interpretation of semantic image concepts, we use image blobs to capture sufficient semantics of image contents [7]. The visual features for image content representation include 1-dimensional coverage ratio (i.e. density ratio) for a coarse shape representation, 7-dimensional LUV dominant colors and color variances, 14-dimensional Tamura texture, and 14-dimensional wavelet texture features.

To enable more effective image classification, concept hierarchy is used for semantic image concept organization and hierarchical classifier training. The concept hierarchy defines the basic vocabulary of the semantic image concepts and their contextual and logical relationships, where the concept nodes on the first level of the concept hierarchy are named as atomic image concepts that can be interpreted.
In our experiments, where the first level represents the underlying atomic image concepts, effectively by using the relevant image blobs and their visual features. One example of the concept hierarchy, that is used in our experiments for a specific domain of natural images, is given in Fig. 1. To achieve more effective image classification, we use hierarchical mixture model to model the semantic image concepts and interpret their inter-level contextual relationships.

As shown in Fig 1, one certain atomic image concept $C_j$ at the first level of the concept hierarchy can be interpreted by using a finite mixture model (FMM) to approximate the class distribution of the relevant image blobs:

$$P(X, C_j, \Theta_{c_j}) = \sum_{l=1}^{\kappa_j} P(X|C_j, \theta_l) \omega_l$$  \hspace{1cm} (1)$$

where $\sum_{l=1}^{\kappa_j} \omega_l = 1$, $\Theta_{c_j} = \{\kappa_j, \omega_{c_j}, \theta_{c_j}\}$ is the parameter set for model structure, weights and model parameters, $P(X|C_j, \theta_l)$ is the $l$th mixture component to characterize the class distribution for the $l$th type of image blobs, $\kappa_j$ is the model structure (i.e., the optimal number of mixture components), $\omega_{c_j} = \{\omega_1, \cdots, \omega_{\kappa_j}\}$ is the weight set for $\kappa_j$ mixture components, $\omega_l$ is the relative weight for the $l$th mixture component to characterize the relative importance of the $l$th type of image blobs for accurately interpreting the given atomic image concept $C_j$, $\theta_{c_j} = \{\theta_1, \cdots, \theta_{\kappa_j}\}$ is the set of model parameters, $\theta_l$ is the model parameter for the $l$th mixture component, $X$ is the $n$-dimensional visual features for representing the relevant image blobs.

The second level of classification is to model the semantic image concepts at the second level of the concept hierarchy by using the relevant atomic image concepts. Thus, the finite mixture model for the second-level semantic image concept $C_i$ can be approximated by integrating the underlying class distribution for the relevant sibling atomic image concepts that take $C_i$ as their parent node:

$$P(X, C_i, \Theta_{c_i}) = \sum_{j=1}^{\kappa_i} \omega_{c_j} \sum_{l=1}^{\kappa_j} P(X|C_j, \theta_l) \omega_{s_l}$$  \hspace{1cm} (2)$$

where $\sum_{j=1}^{\kappa_i} \omega_{c_j} = 1$, $\Theta_{c_i} = \{\kappa_i, \omega_{c_i}, \theta_{c_i}\}$ is the parameter set for model structure, weights and model parameters, and $\omega_{c_i}$ is the set of weights for the gate networks that are used to define the relative importance of the $\kappa_i$ sibling atomic image concepts for accurately interpreting the second-level semantic image concept $C_i$ [11-12].

Through a similar hierarchical approach, one certain semantic image concept $C_k$ at the $n$th level of the concept hierarchy can be interpreted by using a hierarchical mixture model to approximate the class distribution of the sibling lower-level semantic image concepts that take $C_k$ as their parent node:

$$P(X, C_k, \Theta_{c_k}) = \sum_{m=1}^{\kappa_n} \omega_{cm} \cdots \sum_{j=1}^{\kappa_j} \omega_{cj} \sum_{l=1}^{\kappa_l} P(X|C_j, \theta_l) \omega_{s_l}$$  \hspace{1cm} (3)$$

where $\sum_{m=1}^{\kappa_n} \omega_{cm} = 1$, $\Theta_{c_k} = \{\kappa_m, \omega_{cm}, \theta_{cm}\}$ is the parameter set for model structure, weights, and model parameters, and $\omega_{cm}$ is the set of weights for the gate networks that are used to define the relative importance of the sibling lower-level semantic image concepts for accurately interpreting the higher-level semantic image concept $C_k$.

The advantages of our hierarchical framework include: (a) The concept models for interpreting the higher-level semantic image concepts can be adapted by the observations of the relevant atomic image concepts at the first level of the concept hierarchy. In addition, learning the higher-level semantic image concepts hierarchically is able to reduce the size of covariance matrices being inverted. (b) Our hierarchical classifier training framework is able to provide a natural way to achieve discriminative learning of finite mixture models by jointly learning the sibling semantic image concepts under the same parent node, thus our hierarchical classifier training technique is able to achieve higher prediction accuracy.

3 Hierarchical Classifier Training

To accurately learn the hierarchical mixture model for semantic image concept interpretation, the semantic labels for a set of training samples (i.e. image blobs) are manually labeled for each atomic image concept. We use the one-against-all rule to organize the labeled samples $\Omega_{c_j} = \{X_t, C_j(S_t)|t = 1, \cdots, N_L\}$ into: positive and negative samples for one certain atomic image concept $C_j$, $X_t$ is the $n$-dimensional visual features that are used to describe the training sample $S_t$. The unlabeled samples $\Omega_{c_j} = \{X_k, S_k|k = 1, \cdots, N_u\}$ can be used to improve the density estimation by reducing the variance of mixture density and discovering the unknown image context classes. For a certain atomic image concept $C_i$, we then define the mixture training sample set as $\Omega = \Omega_{c_j} \cup \Omega_{c_i}$. 

![Figure 1: The concept hierarchy for organizing natural images](image-url)
The atomic image concepts at the first level of the concept hierarchy have low hypothesis variances and thus they can be accurately interpreted by using the relevant image blobs. Based on this observation, we propose a bottom-up approach for hierarchical classifier training.

3.1 Base-Level Classifier Training

To learn the model-based classifier for the given atomic image concept \( C_j \), maximum likelihood criterion can be used to determine the underlying model parameters. To avoid the overfitting problem [16], a penalty term is added to determine the underlying optimal model structure. The optimal parameters (i.e., model structure, weights, and model parameters) \( \Theta_{c_j} = (\kappa_j, \hat{w}_{c_j}, \hat{\theta}_{c_j}) \) for the given atomic image concept \( C_j \) at the first level of the concept hierarchy are then determined by:

\[
\hat{\Theta}_{c_j} = \arg \max_{\Theta_{c_j}} \{L(C_j, \Theta_{c_j})\} \tag{4}
\]

where \( L(C_j, \Theta_{c_j}) = -\sum_{i \in D_{c_j}} \log P(X_i | C_j, \Theta_{c_j}) + \log p(\Theta_{c_j}) \) is the objective function, \( -\sum_{i \in D_{c_j}} \log P(X_i | C_j, \Theta_{c_j}) \) is the likelihood function, and \( \log p(\Theta_{c_j}) = -\frac{n + \kappa_j + 3}{2} \sum_{i=1}^{\kappa_j} \log \frac{N_\omega}{12} - \frac{\kappa_j}{2} \log \frac{N}{12} - \frac{\kappa_j (N + 1)}{2} \) is the minimum description length (MDL) term to penalize the complex models [16], \( N \) is the total number of training samples, and \( n \) is the dimensions of visual features \( X \).

The estimation of maximum likelihood described in Eq. (4) can be achieved by using the EM algorithm with a pre-defined model structure \( \kappa_j \) [13-15]. However, pre-defining the model structure \( \kappa_j \) is not acceptable for semantic image classification because different atomic image concepts may be related to different numbers and types of various image blobs. Thus, there is an urgent need to develop new techniques that are able to select the optimal model structure automatically. In order to select the optimal model structure \( \kappa_j \) and escape from the local extrema, we propose an adaptive EM algorithm. Our adaptive EM algorithm starts from a large value of \( \kappa_j \) and performs automatic merging, splitting, and elimination of mixture components to select the optimal model structure \( \kappa_j \) and re-organize the distribution of mixture components according to the class distribution of the available training samples.

To select the mixture components for merging, the intra-concept Kullback divergence \( KL(C_j, \theta_i, \theta_k) \) is used to measure the divergence between the \( i \)th mixture component \( P(X|C_j, \theta_i) \) and the \( k \)th mixture component \( P(X|C_j, \theta_k) \) from the same atomic image concept \( C_j \):

\[
KL(C_j, \theta_i, \theta_k) = \int P(X|C_j, \theta_i) \log \frac{P(X|C_j, \theta_i)}{P(X|C_j, \theta_k)} \tag{5}
\]

The strongly overlapping mixture components (i.e. with small value of \( KL(C_j, \theta_i, \theta_k) \)) provide similar densities and overpopulate the relevant image samples. Thus, the overlapped mixture components \( P(X|C_j, \theta_i) \) and \( P(X|C_j, \theta_k) \) can be merged as one single mixture component \( P(X|C_j, \theta_{ik}) \). The local Kullback divergence \( KL(C_j, \theta_{ik}) \) is used to measure the local divergence between the merged mixture component \( P(X|C_j, \theta_{ik}) \) and the local sample density \( P(X|\theta_{ik}) \). To select the best candidate for merging, our adaptive EM algorithm tests \( \omega (\kappa_j - 1) \) pairs of mixture components. The mixture component pair with the minimum value of the intra-concept and local Kullback divergences is selected as the candidate for merging.

Two types of mixture components may be split: (a) The elongated mixture components which underpopulate the relevant samples (i.e., characterized by the local Kullback divergence); (b) The tailed mixture components which overlap with the mixture components from other sibling concept models under the same parent node (i.e., characterized by the inter-concept Kullback divergence). To select the mixture component for splitting, two criteria are combined: (1) The local Kullback divergence \( KL(C_j, \theta_i) \) to characterize the divergence between the \( i \)th mixture component \( P(X|C_j, \theta_i) \) and the local sample density \( P(X|\theta_i) \); (2) The inter-concept Kullback divergence \( KL(C_j, C_h, \theta_i, \theta_m) \) to characterize the overlapping between the mixture components \( P(X|C_j, \theta_i) \) and \( P(X|C_h, \theta_m) \) from two sibling atomic image concepts \( C_j \) and \( C_h \).

If one specific mixture component is only supported by few samples, it may be removed from the underlying concept model. To determine the unrepresentative mixture component for elimination, our adaptive EM algorithm uses the local Kullback divergence \( KL(C_j, \theta_i) \) to characterize the representation of the mixture component \( P(X|C_j, \theta_i) \) for the relevant samples. The mixture component with the maximum value of the local Kullback divergence is selected as the candidate for elimination.

To jointly optimize these three operations of merging, splitting and elimination, their probabilities are defined as:

\[
J_m(i, k, \theta_{ik}) = KL(C_j, \theta_{ik}) + \varphi KL(C_j, \theta_i, \theta_k)
\]

\[
J_s(i, m, \theta_i) = \varphi KL(C_j, C_h, \theta_i, \theta_m)
\]

\[
J_e(i, \theta_i) = \frac{\varphi KL(C_j, \theta_i)}{KL(C_j, \theta_i)}
\]

where \( \varphi \) is a normalized factor and it is determined by:

\[
\sum_{i=1}^{\kappa_j} J_s(i, \theta_i) + \sum_{i=1}^{\kappa_j} \sum_{k=i+1}^{\kappa_j} J_m(i, k, \theta_{ik}) + \sum_{i=1}^{\kappa_j} \sum_{m=1}^{\kappa_h} J_e(i, m, \theta_i) = 1
\]

The acceptance probability to prevent poor operation of merging, splitting or elimination is defined by:

\[
P_{accept} = \min \left( \exp \left[ -\frac{|L(C_j, \Theta_1) - L(C_j, \Theta_2)|}{\tau} \right] , 1 \right)
\]

(7)
where $L(C_j, \Theta_1)$ and $L(C_j, \Theta_2)$ are the objective functions for the models $\Theta_1$ and $\Theta_2$ (i.e. before and after performing the merging, splitting or elimination operation) as described in Eq. (4), $\tau$ is a constant that is determined experimentally. In our current experiments, $\tau$ is set as $\tau = 9.8$.

Our adaptive EM algorithm has the following advantages: (a) It does not require a careful initialization of the model structure and model parameters. By starting from a reasonably large number of mixture components, our adaptive EM algorithm is able to automatically select the optimal model structure to capture the essential structure of the image context classes by performing automatic merging, splitting and elimination of mixture components. Thus, it is able to achieve a better approximation of the real class distribution for the given atomic image concept by running the local search from many different starting points. (b) It is able to improve the prediction power of model-based classifiers by integrating the negative samples to enable discriminative learning of finite mixture models via maximizing the margins among the concept models for the sibling atomic image concepts under the same parent node. (c) It is able to escape the local extrema by re-organizing the distribution of mixture components according to the underlying class distribution of the available training samples. (d) By performing automatic merging, splitting, and elimination of mixture components, it is able to support a new framework for classifier combining (see Section 3.3).

\section{3.2 Learning with Unlabeled Samples}

When only a limited number of labeled samples are available for classifier training, it is difficult to select the optimal model structure and estimate the accurate model parameters. In addition, incorporating the outlying unlabeled samples for classifier training may lead to worse performance rather than improvement. Thus, it is very important to develop new techniques able to eliminate the misleading effects of the outlying unlabeled samples.

After the weak classifier for the given atomic image concept $C_j$ is available, the Bayesian framework is used to achieve “soft” classification of unlabeled images. The confidence score for an unlabeled sample (unlabeled image) with the given atomic image concept $C_j$ is defined as:

$$
\psi(X_l, C_j, t) = \frac{\psi_\alpha(X_l, C_j, t) \psi_\beta(X_l, C_j, t)}{\sqrt{\psi_\alpha(X_l, C_j, t) \psi_\beta(X_l, C_j, t)}} \quad (8)
$$

where $\psi_\alpha(X_l, C_j, t)$ is the posterior probability for the unlabeled sample $\{X_l, S_l\}$ with the given atomic image concept $C_j$, $\psi_\beta(X_l, C_j, t) = -\log P(X_l, C_j, \Theta_{c_j})$ is the log-likelihood value of the unlabeled sample $\{X_l, S_l\}$ with the given atomic image concept $C_j$. For one specific unlabeled sample $\{X_l, S_l\}$, its confidence score $\psi(X_l, C_j, t)$ can be used as the criterion to indicate the possibility to be taken as an outlier for the given atomic image concept $C_j$.

Figure 2: The classification results for the semantic image concept “beach” and the relevant image blobs.

In order to eliminate the misleading effects of the outlying unlabeled samples for semi-supervised classifier training, the unlabeled samples are first categorized into two classes according to their confidence scores: (a) certain unlabeled samples with high confidence scores may originate from the known image context classes that have already been used for interpreting the given atomic image concept $C_j$; (b) uncertain unlabeled samples with low confidence scores may originate from new concept, outliers or unknown image context classes that cannot be directly learned from a limited number of available labeled samples.

The certain unlabeled samples can be incorporated to improve the mixture density estimation incrementally (i.e. regular updating of model parameters without changing the model structure) by reducing the density variance. With the updated concept model for the given atomic image concept $C_j$ (i.e. incremental classifier), the confidence scores for some uncertain unlabeled samples may be changed over time when they originate from the unknown image context classes that cannot be interpreted intuitively by a limited number of labeled samples. For the uncertain unlabeled sample, the changing scale of its confidence scores with the given atomic image concept $C_j$ is defined as:

$$
y_l = |\psi(X_l, C_j, t + 1) - \psi(X_l, C_j, t)| \quad (9)
$$

where $y_l \geq 0$, $\psi(X_l, C_j, t)$ and $\psi(X_l, C_j, t+1)$ indicate its confidence scores with the same concept model $C_j$ before and after updating. The uncertain unlabeled samples with a large value of $y_l$ may originate from the unknown image context classes induced by concept drift, and should therefore be used to achieve more accurate image concept interpretation and semi-supervised classifier training. To address the concept drift problem, one or more new mixture components can be added to the residing areas for these uncertain unlabeled samples with a large value of $y_l$ (i.e. birth). On the other hand, the outlying unlabeled samples with the $y_l$
value close to zero may originate from new concept or outliers. To eliminate the misleading effects of the outlying unlabeled samples, a penalty term $\gamma_l$ is defined as:

$$
\gamma_l = \begin{cases} 
1, & \text{certain unlabeled samples} \\
\frac{e^{y_l}-e^{-y_l}}{e^{y_l}+e^{-y_l}}, & \text{uncertain unlabeled samples}
\end{cases}
$$

(10)

where $0 \leq \gamma_l \leq 1$, $\gamma_l = 0$ if $y_l = 0$. Thus, the penalty term $\gamma_l$ can provide an effective solution to select the informative unlabeled samples for semi-supervised classifier training.

By incorporating only the certain unlabeled samples and the informative unlabeled samples for semi-supervised classifier training, our proposed algorithm has the following advantages: (a) It is able to discover the unknown image context classes and select the optimal model structure for achieving more accurate concept interpretation (i.e., using the birth operation to add new mixture components). (b) It is able to eliminate the misleading effects of the outlying unlabeled samples by using a novel penalization framework. Thus, our new framework is able to improve the classification performance significantly.

### 3.3 Higher-Level Classifier Training

Given the model structures and model parameters for the sibling atomic image concepts at the first level of the concept hierarchy, the next step is to infer the model structure and model parameters for their parent node at the second level of the concept hierarchy (i.e., one certain second-level semantic image concept $C_i$).

Because the model structures and the model parameters for the sibling atomic image concepts have already been obtained, we develop a simple but effective solution to determine the underlying finite mixture model for accurately interpreting their parent node at the second level of the concept hierarchy. Our framework takes the following steps:

(a) The mixture components from the $\kappa_i$ sibling atomic image concepts are combined to achieve a better approximation of the underlying class distribution of their parent node at the second level of the concept hierarchy. In addition, the training samples for these sibling atomic image concepts are also combined as the joint training samples for their parent node $C_i$. (b) Based on the available finite mixture models for interpreting the $\kappa_i$ sibling atomic image concepts, our adaptive EM algorithm is used to select the optimal model structure and estimate the accurate model parameters for the given second-level semantic image concept $C_i$ by performing automatic merging, splitting, and elimination of mixture components. (c) The mixture components with less prediction power on the joint training samples collected from these sibling atomic image concepts are eliminated. The overlapped mixture components from different sibling atomic image concepts are merged into a single mixture component. The elongated mixture components that underpopulate the joint training samples are split into multiple representative mixture components.

If one mixture component, $P(X|C_j, \theta_m)$, is removed, the concept model for accurately interpreting the given second-level semantic image concept $C_i$ is then refined as:

$$
P(X, C_i, \Theta_{c_i}) = \frac{1}{1 - \omega_m} \sum_{l=1}^{\kappa-1} P(X|C_j, \theta_l)\omega_l, \ m \neq l
$$

where $\kappa = \sum_{j=1}^{\kappa_i} \kappa_j$ is the total number of the mixture components for the second-level semantic image concept $C_i$.

If two mixture components $P(X|C_j, \theta_m)$ and $P(X|C_h, \theta_l)$ from two sibling atomic image concepts $C_j$ and $C_h$ are merged as a single mixture component $P(X|C_j, \theta_{ml})$, the concept model for accurately interpreting the second-level semantic image concept $C_i$ is refined as:

$$
P(X, C_i, \Theta_{c_i}) = \sum_{h=1}^{\kappa-2} P(X|C_j, \theta_h)\omega_h + P(X|C_j, \theta_{ml})\omega_{ml}
$$

where $\omega_{ml}$ is the weight parameter for the merged mixture component $P(X|C_j, \theta_{ml})$.

If one mixture component, $P(X|C_j, \theta_h)$, is split into two new mixture components, $P(X|C_j, \theta_r)$ and $P(X|C_j, \theta_t)$, the underlying concept model for accurately interpreting the second-level semantic image concept $C_i$ is refined as:

$$
P(X, C_i, \Theta_{c_i}) = \sum_{h=1}^{\kappa-1} P(X|C_j, \theta_h)\omega_h + P(X|C_j, \theta_r)\omega_r + P(X|C_j, \theta_t)\omega_t
$$

After the finite mixture models for the sibling second-level semantic image concepts are available, they can further be integrated to obtain the finite mixture model for their parent.
node at the third level of the concept hierarchy. Through a hierarchical approach, the hierarchical mixture models for accurately interpreting the higher-level semantic image concepts can be obtained effectively.

After the classifiers are available, they are used to classify the images into different semantic image concepts hierarchically. Our current experiments focus on generating 15 basic semantic image concepts, such as “beach”, “garden”, “mountain view”, “flower view”, “water way”, “sailing”, and “skiing”, which are widely distributed in natural images. Some semantic image classification results are given in Fig. 2, Fig. 3, and Fig. 4.

4. Algorithm Evaluation

Our experiments are conducted on two image databases: the image database from the Google image search engine and the Corel image database. The image database from Google image search engine consists of 30,000 pictures. The Corel image database includes more than 3,800 pictures consisting of different semantic image concepts. Our works on algorithm evaluation focus on: (a) evaluating the performances of our adaptive EM algorithm with different combinations of merging, splitting and elimination; (b) evaluating the performance of our classifier training technique by using different sizes of unlabeled samples.

The benchmark metric for algorithm evaluation includes precision $\rho$ and recall $\varrho$. They are defined as:

$$\rho = \frac{\vartheta}{\vartheta + \gamma}, \quad \varrho = \frac{\vartheta}{\vartheta + \nu} \quad (11)$$

where $\vartheta$ is the set of true positive samples that are related to the corresponding semantic image concept and are classified correctly, $\gamma$ is the set of true negative samples that are irrelevant to the corresponding semantic image concept and are classified incorrectly, and $\nu$ is the set of false positive samples that are related to the corresponding semantic image concept but are misclassified.

In our adaptive EM algorithm, multiple operations, such as merging, splitting, and elimination, have been integrated to re-organize the distribution of mixture components and select the optimal number of mixture components according to the real class distribution of the training samples. Thus, our adaptive EM algorithm is expected to have better performance than the traditional EM algorithm and its recent variants [13].

In order to evaluate the real benefits of the integration of these three operations (i.e. merging, splitting, and elimination), we have compared the performances of our adaptive EM algorithm with different combinations of these three operations. As shown in Fig. 5, we have compared the performances of the classifiers under different combinations of three operations: $SM + Neg$ represents combining three operations of merging, splitting and elimination of mixture components, $SM$ indicates combining two operations of merging and splitting, $Split$ is for only operation of splitting, $Merge$ is for only one operation of merging. From
Fig. 5, one can find that our adaptive EM algorithm is able to achieve more effective classifier training by performing automatic merging, splitting and elimination.

Given a limited number of labeled samples, we have tested the performance of our classifiers by using different sizes of unlabeled samples for classifier training (i.e., with different size ratios $\lambda = \frac{N_u}{N_L}$ between the unlabeled samples $N_u$ and the labeled samples $N_L$). The average performance differences for some semantic image concepts are given in Fig. 6.

One can find that the unlabeled samples can improve the classifier’s performance significantly when only a limited number of labeled samples are available for classifier training. The reasons are: (a) The certain unlabeled samples, that originate from the existing context classes for concept interpretation, are able to improve the density estimation by reducing the variances of mixture density. (b) The informative unlabeled samples, that originate from the unknown context classes, have the capability to provide additional context knowledge to learn the concept models more accurately (i.e., via birth operation). By modifying the concept models to be more representative for the data resources, the concept models that are learned incrementally are able to determine the classifiers with higher prediction accuracy. (c) The outlying unlabeled samples, that originate from outliers, can be predicted and their misleading effects on classifier training can be eliminated automatically by using a novel penalization framework.

When a limited number of labeled samples are available and more unlabeled samples are involved for semi-supervised classifier training (i.e., $\lambda = \frac{N_u}{N_L}$ becomes bigger), we have also obtained a decrease in the classifier’s performance because large-scale outlying unlabeled samples have dominated the statistical properties of the joint class distribution and misled the classifier.

5. Conclusions

When only a limited number of labeled samples are available for classifier training, we have developed a novel framework to incorporate the unlabeled samples and the concept hierarchy for achieving more accurate model selection and parameter estimation. With a suitable penalization technique, our new framework is able to eliminate the misleading effects of the outlying unlabeled samples for semi-supervised classifier training. By integrating the concept hierarchy for semantic image concept organization and hierarchical classifier training, one key advantage of our new framework is that it is able to enable discriminative learning of finite mixture models by using the negative samples to maximize the margins among the concept models for the sibling semantic image concepts.

References


