Semi Supervised Learning in Wild Faces and Videos

David Rim
david.rim@polymtl.ca
Kamrul Hassan
md-kamrul.hasan@polymtl.ca
Chris Pal
christopher.pal@polymtl.ca

Ecole-Polytechnique Montreal
Montreal, Quebec, CA

Abstract

We propose an approach for improving unconstrained face recognition based on leveraging weakly labeled web videos. It is easy to obtain videos that are likely to contain a face of interest from sites such as YouTube through issuing queries with a person's name; however, many examples of faces not belonging to the person of interest will be present. We propose a new technique capable of learning using weakly or noisily labeled faces obtained in this setting. In particular, we present a novel method for semi-supervised learning using noisy labels which incorporates a margin or null category like property within a fully probabilistic framework. We outline general properties of the approach, showing how the choice of an exponential hyperprior results in an L1 penality which leads to sparse models capable of explicitly accounting for label uncertainty producing state of the art performance. We then illustrate how the margin approach provides robustness and significant performance gains when faces within YouTube search results are combined with the unconstrained face images from the Labeled Faces in the wild dataset.

1 Introduction

Facial recognition research has recently focused on the more difficult task of recognition in unconstrained images – images of faces in commonly occuring conditions. These images are usually produced by consumer digital photographers under conditions of varying illumination and pose, often occluded (*e.g.* with glasses), sometimes heavily compressed or degraded by motion blurring. Despite much work in recent years, facial recognition in these conditions remains a difficult task. In the case of unconstrained facial recognition, using only labeled examples, where the number of labeled examples may only be on the order of hundreds, is a primary source of this difficulty.

Finding sources of unlabeled facial images is not a challenging task. Several large web-based collections such as Flickr and Google Images provide public access to millions of static images. However, while static images may provide a large number of examples, the numbers of such images compare poorly to those available in video. Furthermore, facial images in video have properties which differ from static images – facial images sampled

BMVC 2011 http://dx.doi.org/10.5244/C.25.3

from a video will tend to have similar if not identical backgrounds, common illumination and other common properties which may help to model the manifold-like properties of a single subject's face. At the same time, facial images in video are even less constrained than in static images, making recognition in video a difficult task even with large amounts of labeled static images.

However, we believe that video simply provides more information than static images alone. To that end, we create a large new dataset of facial images sampled from Youtube videos designed to mirror the Labeled Faces in the Wild dataset [9]. Although the database is unlabeled, information related to each video provides context for facial images contained within the video. We use the simple approach of using search queries as a weak label. We then use simple descriptor based methods shown to have good performance in the unconstrained face verification task and combine this feature representation with a novel probablistic model with a low-density separation approach and show that unconstrained facial recognition in both static and video images can be significantly improved using unlabeled data.

1.1 Related Work

Related Work on Facial Recognition The Labeled Faces in the Wild Dataset (LFW), created by Huang *et al.* [9], provides a natural composition of face images. However, the mean number of training images per subject in this dataset is little more than 2, with roughly less than 30% of the data available as use for traditional train and test supervised classification. This makes facial recognition in the traditional sense a difficult task. As such, the stated focus of the LFW database is on the pair matching task or one-example learning, which roughly shares the same objective [9]. Pair matching using the LFW dataset is quite mature, with current accuracy of better than 88% as in [18] and [20].

Pair matching, however, is not an identical problem to the face verification task as coined by Huang et al [9]. In face verification, an algorithm is tasked to label a test image as belonging to one of a set of subjects. Although a pair matching algorithm can be used in a nearest-neighbor fashion, the standard approaches of multi-class classification can also be brought to bear. The work by Wolf *et al.* [25] and [24] are the most representative of this work. [24] specifically addresses the question of how well descriptor-based methods often used in verification tasks in object recognition work for pair matching. Wolf *et al.* note that for classes with a relatively large number of training examples, (greater than 10), resulting in a subset of classes, quite good results can be achieved [24]. It therefore appears that the main issue is the number of labeled positives in the LFW dataset. Naturally, this raises the question of whether semi-supervised approaches can be utilized instead of labeling many images by hand.

Semi-Supervised Learning Margin-based or margin-like properties have been presented before in the context of semi-supervised learning *i.e.*. low density separation. Entropy regularization [7] finds a classifier in which the classes of the unlabeled data are maximally separated according to the entropy of the conditional distribution. Assuming absence of a weak label, and not using a null category, the resulting objective function is similar in spirit. Transductive support vector machines (SVM) [11], [12] find a decision boundary consistent with the labeled data which maximally separates both the labeled and test (unlabeled) data. Meanwhile, incorporating prior knowledge in a manner similar to our method using a weighted margin SVM and confidence values obtained from unlabeled data is presented in [26]. The null category noise model is presented in the context of Gaussian processes to produce decision boundaries in regions of low density in an attempt to produce a margin-like effect in a fully Bayesian framework [15]. This paper attempts to present a similar low-density separation method which softens the hard constraints used in the the transductive SVM and the null-category noise model, and also extends this framework to use weak labels.

The use of weak labels of this nature is closely related to the Generalized Expectation Criteria of McCallum, Mann and Druck [17], [4], and to the Expectation Regularization of Mann and McCallum [16]. In the latter case, the algorithm is expected to match the proportion of labels on the unlabeled data in a global sense, using a temperature setting to help enforce "peaky" imputed labels. Here, the weak labelings are local estimates, using a null category to avoid labeling data points which are low confidence. The method most similar to our type of prior information is in Zhu and Ghahramani's graph based algorithm which incorporates class prior information [27], however, there the emphasis is on the graph method and no attempt is made to make the class priors robust.

2 Model

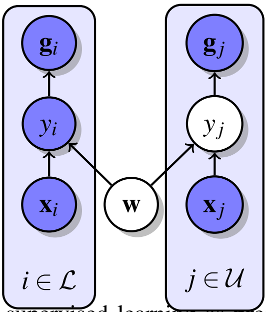


Figure 1: Graphical model of semi-supervised learning as presented in [15], $\mathcal{U}$ is the index set of unlabeled examples, and $\mathcal{L}$ is the index set of the labeled examples.

In this framework, a dataset $\mathcal{D}$ is composed of sets of variables $\mathcal{D} = \{(\mathbf{x}_n, y_n, \mathbf{g}_n)\}_{n=1,2,\ldots,N}$, with $\mathbf{x} \in \mathbb{R}^d$, $y \in \{-1,1\}$, and $\mathbf{g} \in \{-1,1\}$, where the conditional dependencies are specified as in the model shown in Figure (1). That is, the data have labels, y, and the task is to output a classifier parameterized by the random vector $\mathbf{w}$, which, when combined with an observed $\mathbf{x}$, maps to the correct label y. In this set up not every example has a corresponding label, giving the general framework of binary classification in the presence of missing values. Let $\mathcal{L}$ be an index set for the labeled data $\mathcal{D}_i$, and $\mathcal{U}$ be an index set for the unlabeled data $\mathcal{D} \setminus \mathcal{L}$ The dependencies shown in the model above allows us to use the additional observed values $\mathbf{g}$ to train a discriminative classifier using the unlabeled data. We note that because of the d-separation of $\mathbf{w}$ and $\mathbf{g}_i$ by y_i, it is unnecessary to learn a maximum likelihood solution for $\mathbf{w}$ if the labels are observed. However, if the label, y_j, is missing, then observing $\mathbf{g}_j$ makes $\mathbf{w}$ and $\mathbf{x}_j$ conditionally dependent, having the common observed ancestor, seen in Figure (1), so that unlabeled data are useful. If the relationship between y_j and g_j is tightly coupled, then g_j may be thought of as a noisy label. The general idea is that g_j be a variable which is far easier to obtain than the true label, y_j. The above model can be learned using a variational approach [13], in which the objective is to maximize the log marginal distribution

$$\log p(\mathcal{D}) = \int_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathcal{D},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} + KL(q||p) = \mathcal{L}(q) + KL(q||p) \tag{1}$$

where $\mathbf{Z}$ are the latent variables. Treating the parameter $\mathbf{w}$ as a hidden as well, $\mathbf{Z} = \{\mathbf{w}, (y_j)_{\{j\in U\}}\}$ and $\mathcal{D} = \{(y_i)_{i\in L}, \mathbf{g}, \mathbf{X}\}$. Since $KL(q||p) \geq 0$, $\mathcal{L}(q) \leq \log p(\mathcal{D})$ and maximizing $L(q)$ in (1) with respect to the distribution $q(\mathbf{Z})$ is equivalent to minimizing the KL divergence $KL(q||p)$. As is typically the case in variational inference [1], in what Gharamani and Beal term the variational approach, [6], we let $q(\mathbf{Z})$ factorize with respect to each y_j for $j \in U$, and $\mathbf{w}$ as disjoint sets. That is, $q(\mathbf{Z}) = q(\mathbf{w})\prod_j q(y_j)$. Expanding the terms according to the graphical model we have above, we see that the objective is to maximize $\mathcal{L}(q)$, which decomposes into expectations over the unlabeled and labeled data, so that we can expand the above to give

$$\begin{aligned}
\mathcal{L}(q) &= \mathcal{L}_{\text{labeled}} + \mathcal{L}_{\text{unlabeled}} && (2)\\
&= \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{\prod_{i\in L} p(y_i|\mathbf{x}_i,\mathbf{w})p(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} \\
&+ \int_{\mathbf{w}} \sum_{y_{L+1}} \sum_{y_{L+2}} \dots \sum_{y_U} q(\mathbf{w}) \prod_j q(y_j) \log \frac{\prod_{j\in U} p(g_j|y_j)p(y_j|\mathbf{x}_j,\mathbf{w})}{\prod_j q(y_j)} d\mathbf{w}. && (3)
\end{aligned}$$

The first term can be seen as the expectation of $\mathbb{E}_{\mathbf{w}}[\log p(D_l)] + \mathbb{H}[q(\mathbf{w})]$. The second term encapsulates $|U|$ KL divergence-like terms. To make this clearer, we can rewrite the second term as

$$\mathcal{L}_{\text{unlabeled}} = \sum_j \sum_{y_j} q(y_j) \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(g_j|y_j)p(y_j|\mathbf{x}_j,\mathbf{w})}{q(y_j)} d\mathbf{w} \tag{4}$$

In the notation of [17], [4], this is very close to the Generalized Expectation Criteria (GEC)

$$-D(\tilde{g}_{xy}||E_{\{\mathbf{Z}\setminus y_j\}}[p(y_j|x_j,\mathbf{Z})G(x_j,y_j)]), \tag{5}$$

where $G(x,y) = p(g_j|y_j)$, $p(y_j|x_j,\mathbf{Z}) = p(y_j|\mathbf{x}_j,\mathbf{w})$, and $\tilde{g}_{xy}$ is the distribution $q(y_j)$. However, here, we maximize $\mathbb{E}_{\mathbf{w}}[\log p(g_j|y_j)p(y_j|\mathbf{x}_j,\mathbf{w})]$, rather than $\log \mathbb{E}_{\mathbf{w}}[p(g_j|y_j)p(y_j|\mathbf{x}_j,\mathbf{w})]$, which is an upper bound. The unlabeled data acts as a regularization penalty on a quantity close to the divergence between the variational distribution $\prod_j q(y_j)$ and the expected distribution of the data under $q(\mathbf{w})$. The extension of these quantities to more general constraint functions $G(x_j,y_j)$ is the primary contribution of the GEC. However, we note that if $G(x_j,y_j)$ is a probability distribution, it can also be interpreted as a variational distribution.

Null Category We treat the null category as in [15] as an additional Y-label which takes the value zero. This makes the classification, multiclass, where $Y \in \{1,-1,0\}$. If $p(Y=0) = 0$ when the label is unobserved, the null class will provide a probabilistic "margin", pushing the decision boundary away from unlabeled examples. We extend this intuition by relaxing this restriction and allowing $p(Y=0)$ to be non-zero, yielding a soft margin.

Here, we assume that g is an informative weak label – the unlabeled data is labeled with the class having the larger proportion. We also make the assumption that the unlabeled data will be skewed in favor of the positive class. It is possible to model for varying weak labels as well, the assumption is made here for simplicity. Treating the distribution $p(g_j|y_j)$ as a multinomial, let

$$p(g_j = 1|y_j) = \begin{cases} \mu^+ & \text{if } y_j = 1, \\ \mu^- & \text{if } y_j = -1, \\ \mu^0 = 1 - \mu^+ - \mu^- & \text{otherwise, if } y_j = 0 \end{cases} \tag{6}$$

Let $\mu = (\mu^+, \mu^-)$ be a vector of weights which serve as a bias for each class, much like class priors. Assuming that the model is symmetric, that is, $p(g_j = -1|y_j = -1) = p(g_j = 1|y_j = 1)$, the above specifies a proper distribution. Furthermore, this distribution is normalized by $Z^{-1} = (1 + \exp(\mu^+) + \exp(\mu^-))^{-1}$. Letting $p(y_n|\mathbf{x}_n)$ also be a soft max function of $\mathbf{w}$, (multinomial logistic regression) with $\mathbf{z}_n$ a binary vector of indicator functions, $\mathbf{z}_n = (\mathbf{1}_{\{y_n=1\}}, \mathbf{1}_{\{y_n=-1\}}, \mathbf{1}_{\{y_n=0\}})^T$, then the joint distribution can be expressed as

$$p(D, \{y\}) = \prod_{i\in L} \prod_{c} \frac{\exp(\mathbf{w}_c^T \phi(\mathbf{x}_i))^{z_{ic}}}{C_i} \prod_{j\in U} \prod_{c} \frac{\exp(\mathbf{w}_c^T \phi(\mathbf{x}_j) + z_{jc}\mu_c)^{z_{jc}}}{C_j Z_j} \tag{7}$$

where c indexes the class labels $\{1, -1, 0\}$. Here, C_i^{-1} normalizes $\prod_c \exp(\mathbf{w}_c^T \phi(\mathbf{x}_i))^{z_{ic}}$. In the case of the labeled data, $C_i = \sum_{c\neq 3} \exp(\mathbf{w}_c^T \phi(\mathbf{x}_i))$, as $p(y_i = 0|\mathbf{x}, \mathbf{w}) = 0$, by assumption. ϕ is a transformation function, and in the case of kernel logistic regression, a vector of kernel basis functions, that is, $\phi(x_i) = (K(x_i, x_j))_{\{j=1,2,\ldots,N\}}$. Again, $\mathbf{1}_{\{y_i=0\}} = 0$ for all $i \in L$.

Furthermore, we model the distribution of w_{cn} with a zero-mean Gaussian prior, as in the Relevance Vector Machine (RVM) [21], except that instead of the the gamma hyperprior, as used in Bishop and Tipping's variational approach [2], we use an Exponential hyperprior for the variance parameterized by $\frac{\gamma}{2}$. This can be shown to result in an Exponential prior, which approximates an L_1 penalty, as opposed to an L_2 penalty which will not generally yield sparse $\mathbf{w}$, and provides more shrinkage than the L_0 approximating penalty of the RVM [14]. More formally, we let $w_{nc} \sim N(0|\alpha_{nc})$, and $\alpha_{nc} \sim \text{Exp}(\frac{\gamma}{2})$, for all n and all classes c. We note that the form of the objective results in an RVM-like formulation of a semi-supervised learning algorithm. The inclusion of the Exponential hyperprior results in intractable expectations in the variable $\mathbf{w}$; we use the Exponential hyperprior because it was observed to lead to accurate and sparse classifiers. The expanded joint likelihood can then be written as

$$\prod_{nc} N(w_{nc}|0, \alpha_{nc}) \exp(-\frac{\gamma}{2}\alpha_{nc}) \prod_{i\in L} \left(\frac{\prod_c \exp(\mathbf{w}_c^T \phi(\mathbf{x}_i))}{C_i} \right)^{z_{ic}} \prod_{j\in U} \left(\frac{\exp(\mathbf{w}_c^T \phi(\mathbf{x}_j) + z_{jc}\mu_c)}{C_j Z_j} \right)^{z_{jc}}. \tag{8}$$

As $\alpha = (\alpha_{nc})_{n=1,2,\ldots,N,c=1,2,3}$ represents hidden variables as well, we augment (2) with an additional variational distribution $q(\alpha) = \prod q(\alpha_{nc})$. Now, plugging (8) into (2), we obtain the objective function as $\mathcal{L}(q)$. However, for reasons mentioned above, we treat $\mathbf{w}$ as a parameter to $\mathcal{L}(q)$, that is $\mathcal{L}(q,\mathbf{w}) =$

$$\begin{aligned}
&\sum_i \sum_c z_{ic} \mathbf{w}_c^T \phi(\mathbf{x}_i) - \sum_i \log C_i \\
&+ \int q(\alpha) \Big(-\frac{1}{2} \sum_{nc} \log(\alpha_{nc}) - \sum_{nc} \frac{1}{2} \alpha_{nc}^{-1} w_{nc}^2 - \frac{\gamma}{2} \sum_{nc} \alpha_{nc} \Big) d\alpha \\
&+ \sum_{j=|L|+1}^{|U|} \sum_{y_j} q(y_j) \Big(\sum_c z_{jc} \mathbf{w}_c^T \phi(\mathbf{x}_j) + z_{jc} \mu_c - \sum_{k=|L|+1}^{|U|} \log(C_k Z_k) \Big) \\
&- \sum_{j=|L|+1}^{|U|} \sum_{y_j} q(y_j) \log q(y_j) - \int q(\alpha) \log q(\alpha) d\alpha \qquad (9)
\end{aligned}$$

The first term can be thought of as the log likelihood of the labeled data, the second term as a prior or regularization term. The third term is the effect of the unlabeled data, or as an additional regularization term which takes into account the prior likelihood of the classes. The final two terms penalize the entropy of the variational distributions.

Expectation The distribution $q(y_j)$ can be found by noting that the bound is tightest when

$$q(y_j) \propto \prod_c \exp(\mathbf{w}_c^T \phi(\mathbf{x}_j) + z_{jc} \mu_c)^{z_{jc}} \qquad (10)$$

So that the factorized distribution for each y_j unlabeled is given by

$$q(y_j) = \frac{\exp(\mathbf{w}_c^T \phi(\mathbf{x}_j) + z_{jc} \mu_c)^{z_{jc}}}{\sum_c \exp(\mathbf{w}_c^T \phi(\mathbf{x}_j) + \mu_c)} \qquad (11)$$

As the expectation of an indicator function is a probability, we have that $\mathbb{E}_{\mathbf{Z} \setminus y_j}[z_j] = (q(y_j = 1), q(y_j = -1), q(y_j = 0))$. Meanwhile, the distribution of α_n is

$$q(\alpha_{nc}) \propto \alpha_{nc}^{-\frac{1}{2}} \exp(-\frac{1}{2}(\gamma \alpha_{nc} + w_{nc}^2 \alpha_{nc}^{-1})) \qquad (12)$$

Changing variables, let $\tau_{nc} = \alpha_{nc}^{-1}$, resulting in the distribution

$$q(\tau_{nc}) \propto \tau_{nc}^{-\frac{3}{2}} \exp(-\frac{1}{2}(\gamma \tau_{nc}^{-1} + w_{nc}^2 \tau_{nc})), \qquad (13)$$

which is an inverse Gaussian distribution with mean $\frac{\sqrt{\gamma}}{|w_{nc}|}$, and shape parameter $\sqrt{\gamma}|w_{nc}|$ [3].

Maximization As noted, the choice of hyperprior leads to intractable expectations, and as such, we estimate a mode of $p(\mathbf{w})$. Using $\langle . \rangle$ to denote an expectation, the gradient of the marginal distribution $\log p(\mathcal{D})$ with respect to the vector $\mathbf{w}_k$, can be written as

$$\nabla(\mathbf{w}_k)\mathcal{L}(q,\mathbf{w}) = \sum_n \langle z_{nk} \rangle \phi(\mathbf{x}_n) - \sum_n \frac{\exp(\mathbf{w}_k \phi(\mathbf{x}_n))}{C_n} \phi(\mathbf{x}_n) - \langle \mathrm{T}_k \rangle \mathbf{w}_k, \qquad (14)$$

where T_k is the n by n diagonal matrix of expectations $\mathrm{diag}(\langle \tau_{nk} \rangle)$. The gradient is a standard kernel multinomial logistic regression MAP optimization except that the indicators for the unlabeled data are replaced by expectations, in this case, $q(y_j)$ (with some abuse of notation, we let $\langle z_i \rangle$ for $i \in L$ be z_i.) We use an IRLS method with the above to maximize the parameter $\mathbf{w}$. To monitor convergence, we track the increase in the objective function (9), at each iteration until convergence.

	SVM		KLR + Noisy Labels	
Held out	Accuracy	SE	Accuracy	SE
All but 1	0.236	0.014	**0.603**	0.012
0.9000	0.449	0.012	**0.610**	0.004
0.7500	0.604	0.013	**0.651**	0.004
0.5000	0.709	0.013	**0.754**	0.001

Table 1: Accuracy using different proportions of labeled and unlabeled data using a known weak label accuracy parameter. The held out column presents the percentage of data used as unlabeled data. For our method, γ and μ^0 were set to 1 and .50, respectively, *i.e.*. $\mu^+ = .75(1-\mu^0)$.

Testing As the algorithm may not be supported with weak labels at test time, a prediction is based on assuming that a test example is not unlabeled. The probability is then given by the two vectors of weights $\mathbf{w}^+$ and $\mathbf{w}^-$, which can be combined into a binary classifier $p(y=1|\mathbf{x}) = \frac{1}{1+\exp(-\hat{\mathbf{w}}^T\phi(\mathbf{x}))}$, where $\hat{\mathbf{w}} = \mathbf{w}_+ - \mathbf{w}_-$, as $p(y=0|\mathbf{x}) = 0$, by assumption. Note that instead of integrating over parameters as would be the case in a fully Bayesian procedure, we use the MAP estimate $\mathbf{w}$ recovered from the above algorithm for computational reasons.

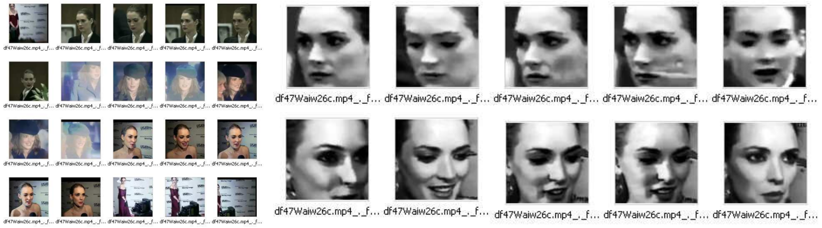
3 Experiments

Labeled Faces in the Wild To test the accuracy of our method under a known weak label accuracy level, we created a subset of the LFW dataset using the 50 subjects having the most labeled examples, yielding 2733 total labeled examples. As in [24], we combine this set with 4000 negatives drawn from the remaining subjects having as least 3 examples, a subset of 6733 examples from LFW. We tested the algorithm using varying amounts of labeled examples by holding out a percentage of each subject's images, and creating a synthetic unlabeled set with 75% accurate weak labels.

The data in Table (1) shows the results of using a linear kernel with the LBP features as input for each held-out regime, with parameters determined by cross-validation. The results indicate that a significant increase in accuracy is possible using unlabeled data, especially in the case where only a very small number of positive labels are available. The additional improvement decreases as the ratio of labeled training examples to unlabeled examples increases. However, at a known level of weak label accuracy, the improvement is still significant, indicating that additional labeled data may be approximated by large amount of weakly labeled data. Note that the best accuracy was not at a value of $\mu^0 = 0$, as would be the case with hard-constraints.

Youtube Video Using the subset of subjects created in the above section, we used the corresponding names for these individuals to download a set of videos from YouTube. In order to obtain videos likely to contain additional faces of interest the full name of each subject was issued as a query, *e.g*. "George W Bush". We downloaded a maximum of thirty videos ranked by YouTube's search for each subject. This procedure resulted in 1277 videos comprising 28.8GB of data.

To avoid returning near duplicates, we collect faces contained in key frames using the ffmpeg [5] and MEncoder tools, employing the OpenCV implementation of the Viola-Jones face dectector to detect and localize faces [8], [22]. To retain high resolution in the resulting face images, we search for faces sized at least the maximum of 45% of the height of the video or 109 × 109. Each of the positive face detection is cropped and rescaled following the identical procedure as in LFW[9]. After processing these 1277 videos, a total number of $42,255$ faces were extracted. False negative face detections were filtered out by running a eye-pair detector on the extracted faces, again provided by OpenCV [8], resulting in $25,726$ faces. We then aligned the face images extracted from the videos using the funneling methodology utilized by Huang *et al.* [10] for the LFW database. Following [23], the images were then cropped to a 110 × 115 window around its center and converted to grayscale. An adaptive noise removal filter (*wiener2*) was used for noise removal and the denoised images were normalized such that 1% of the pixels at the both the highest and lowest ends are saturated. For each of the preprocessed quality faces, we compute the LBP, [19], FPLBP and TPLBP as described in [24], and concatenate them into a single feature vector. The pipeline is described visually in Figure (2).



(a) A sequence of faces detected by the V. and Jones face detector

(b) The sequence after eye-pair filtering, alignment and cropping

Figure 2: The pipeline output for one of Winona Ryder's videos

We test the model using a small amount of labels for each video to estimate the weak-label parameters μ, labeling 4,473 random images of the available 20,765 facial images by random sampling. The sampling procedure provides 2,369 additional positive examples for 50 subjects. We use the label proportions of the sampled data and regress these estimates to the global mean of 53%. We note that this effectively provides a prior to the binomial distribution centered on the mean. We combine these estimates with a relatively large margin size of 50%, to yield a multinomial distribution. To test our procedure in this case, we trained a linear SVM on features derived only from the labeled Youtube face images and evaluated on testing samples of 100 images comprised of two images for each subject. The training and testing were again run 10 times across different train/test splits, resulting in an accuracy of 81.6%. The same experiment was repeated using a Gaussian kernel resulting in a lower accuracy of 78.0%. The γ parameter in the Gaussian kernel was found by cross-validation over a grid. We then use our method and the unlabeled facial images for each subject and a null category probability μ^0, of 75%, yielding accuracies of 75.6% and 85.8% respectively for the linear and Gaussian kernels, respectively. The results seem to indicate that training with a lower number of labeled examples is prone to over-fitting, requiring the use of a more restricted class of classifier, *i.e.*. linear classifiers. However, in the presence of a higher number of examples in the semi-supervised case, the linear classifer under-fits, and the more flexible Gaussian classifier is regularized appropriately by the unlabeled data. Table (2) summarizes these results. We then hypothesized that adding the Youtube facial images to the LFW dataset would yield increased performance using our model based on the larger number of positive examples in the unlabeled set, but would require a more flexible

	SVM		KLR + Noisy Labels	
Kernel	Accuracy	SE	Accuracy	SE
Linear	0.816	0.019	0.770	0.015
Gaussian	0.780	0.009	**0.858**	0.012

Table 2: Accuracy on the Facial Images Recovered from Youtube (tested on Youtube data)

	SVM		KLR + Noisy Labels	
Kernel	Accuracy	SE	Accuracy	SE
Linear	0.781	0.007	0.646	0.016
Gaussian	0.763	0.035	**0.818**	0.013

Table 3: Accuracy on the LFW augmented with Youtube Data (tested on LFW examples)

classifier. The baseline experiment, as described previously is a linear classifier trained using all but 2 of the available labeled LFW images, which were used for testing. These were combined with the labeled examples from the Youtube data. Both a linear and Gaussian SVM were trained using only the labeled data, and then our method was used by adding the remaining unlabeled Youtube images. Again, this set of experiments was repeated 10 times over differing train/test splits. Similar to the Youtube only experiments, it is apparent that the more flexible Gaussian kernel is preferable in the semi-supervised case. The linear kernel case indicates under-fitting. The results are presented in Figure (3). We repeated the above experiment on a final task, to attempt to combine static images with unlabeled video images in order to better classify the video images, *i.e.*. we train using the same dataset but test on video images. The same behavior is again apparent, with the unlabeled data helping to create a better classifier for the video, as shown in Table (4). We note that we prohibited training on images drawn from videos present in the test set.

4 Conclusions and Future Work

Although the gain in some of the experiments described in the previous section is moderate, we believe this is largely due to domain differences. Facial images drawn from the LFW dataset are derived from static news images which contain mostly rectified faces usually centered in the photo.The facial images drawn from Youtube exhibit quite a large amount of variability due to differences in environment and compression artifacts. A direction for future work would include attempting to account for the differences. Computationally, we recognize the limitations of kernel methods with non-convex objectives, specifically the $O(n^3)$ overhead of generating kernel matrices. Much of the computational overhead, however, can be parallelized using modern GPGPU-based methods, including construction of Hessian matrices and matrix inversion required for the IRLS updates. The optimization converges very

	SVM		KLR + Noisy Labels	
Kernel	Accuracy	SE	Accuracy	SE
Linear	0.823	0.004	0.710	0.018
Gaussian	0.775	0.010	**0.861**	0.005

Table 4: Accuracy on the LFW augmented with Youtube Data (tested on Youtube examples)

quickly in IRLS updates, usually fewer than four iterations. We further note that the use of a subject noise level does not preclude using per example noise estimates. That is, we have used a per subject estimate of the accuracy of the weak label. We believe that using a per example noise estimate determined by using text data derived from video meta-data, such as a video description and date, as well as other features of the video should yield better results. We have explored the idea of using these alternative sources as a way of incorporating more complex prior data in the future.

However, we have shown the overall effectiveness of using weak labels derived from video data combined with existing facial images to aid learning tasks. Noisy labels are readily available in video for faces using this method. Adding a null-category with soft constraints produces better results than using the labeled data alone. In all cases, the use of unlabeled data is able to improve the final classifier. The model is fully probabilistic and gives greater flexibility including the use of non-standard priors, most specifically sparsity-inducing priors. Going beyond experiments with a known noise level, using a realistic dataset and an estimate of the noise yielded improvement as well. On a realistic task of interest, improving a static face classifier using readily available video images, the method also produces good results.

References

[1] H. Attias. A variational Bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2):209–215, 2000.

[2] C.M. Bishop and M.E. Tipping. Variational relevance vector machines. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53, 2000.

[3] R.S. Chhikara and L. Folks. *The inverse Gaussian distribution: theory, methodology, and applications*. CRC, 1989.

[4] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *SIGIR*, pages 595–602. ACM, 2008.

[5] ffmpeg. ffmpeg. http://www.ffmpeg.org.

[6] Zoubin Ghahramani and Matthew Beal. Graphical models and variational methods. In Manfred Opper and David Saad, editors, *Advanced mean field methods: theory and practice*. MIT Press, 2001.

[7] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. *NIPS*, 17:529–536, 2004.

[8] Modesto Castrillon-Santana Hannes Kruppa and Bernt Schiele. Fast and robust face finding via local context. In *IEEE Workshop on Visual Surveillance and PETS*, October 2003.

[9] G. B Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *UMass, Amherst, TR 07*, 49:1, 2007. URL http://vis-www.cs.umass.edu/papers/eccv2008-lfw.pdf.

[10] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. of ICCV*, pages 153–160, Rio de Janeiro, Brazil, 2007.

[11] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209. Citeseer, 1999.

[12] T. Joachims. Learning to classify text using support vector machines: Methods, theory, and algorithms. *Computational Linguistics*, 29(4):656–664, 2002.

[13] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[14] B. Krishnapuram, L. Carin, M.A.T. Figueiredo, and A.J. Hartemink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):957–968, june 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.127.

[15] N.D. Lawrence and M.I. Jordan. Semi-supervised learning via Gaussian processes. *NIPS*, 17:753–760, 2005.

[16] G.S. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, page 600. ACM, 2007.

[17] A. McCallum, G. Mann, and G. Druck. Generalized expectation criteria. *UMass, Amherst, TR*, 2007.

[18] Hieu V. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[19] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, pages 971–987, 2002. ISSN 0162-8828.

[20] Nicolas Pinto and David Cox. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2011.

[21] ME Tipping. The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, 12, 2000.

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, pages 511–518, 2001.

[23] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*. Citeseer, 2008.

[24] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008. URL http://www.cs.tau.ac.il/~wolf/papers/patchlbp.pdf.

[25] L. Wolf, T. Hassner, and Y. Taigman. The One-Shot similarity kernel. In *Proc. of ICCV*, September 2009. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5459323.

[26] X. Wu and R. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *KDD*, page 333. ACM, 2004.

[27] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, volume 20, page 912, 2003.