SEMI-SUPERVISED REGRESSION WITH TEMPORAL IMAGE SEQUENCES

Ling Xie, Miguel Á. Carreira-Perpiñán and Shawn Newsam

Electrical Engineering and Computer Science, University of California at Merced

ABSTRACT

We consider a semi-supervised regression setting where we have temporal sequences of partially labeled data, under the assumption that the labels should vary slowly along a sequence, but that nearby points in input space may have drastically different labels. The setting is motivated by problems such as determining the time of the day or the level of air visibility given an image of a landscape, which is hard because the time or visibility label is related in a complex way with the pixel values. We propose a regression framework regularized with a graph Laplacian prior, where the graph is given by the sequential information. We show this outperforms graphs learned in an unsupervised way for detecting the rotation of MNIST digits and estimating the time of day an image is captured, and provides modest improvement in the challenging visibility problem.

Index Terms— semi-supervised learning, scene estimation

1. INTRODUCTION

While specialized equipment is now available to measure the attenuation and scattering of light, using low cost commodity digital cameras to estimate atmospheric visibility holds great appeal. The challenge, of course, is to learn the likely non-linear mapping from the complex image space to an index of visibility. Even in a static scene, increased or decreased visibility due to the attenuation and scattering of light is only one of many possible sources for image variation and usually results in subtle differences. While specialized imaging systems for visibility estimation have been proposed [9, 10], there has been little work on using general purpose cameras.

A similar example is estimating the time of the day from an image from a given landscape. Intuitively, the overall illumination, color, the arrangement of shadows, etc. change in a manner that is strongly correlated with the time of the day in a given scene, as shown in [8]. However, there exist many other changes in the image that are uncorrelated with it: moving objects (cars, clouds, planes, etc.) which cause changes in the scene but also in the overall illumination; lights and reflections from objects; etc. With so much variability in the image space, and settings with little labeled data, how can we learn a predictive mapping (regression or classification) that is able to detect the inputs that really matter?

The potential disconnect between image features and environmental conditions presents a particularly difficult challenge for semi-supervised learning techniques where not all of the training data is labeled. Semi-supervised learning techniques typically exploit the topology of the input space to propagate the labels in the training set. However, two images of a static scene could appear very similar—that is, the distance between them in input space could be small—but could correspond to very different environmental conditions. Take, for example, the three images in figure 1. The two images which are only 20 minutes apart are visually less similar than the two images which are twelve hours apart. Propagating labels, here the time of day, between the dawn and dusk images would most certainly result in a less effective learning process.

Fig. 1: Three sample images from the time of day problem.

We introduce a technique by which additional information about the data in the form of temporal priors is shown to be more effective for semi-supervised learning than techniques which exploit the topology of the input space. We first study the technique in a synthetic problem where we want to learn functions that depend on the input space in a complex way, but where the availability of temporal information provides crucial clues to predict the label. We then consider two problems: estimating time of the day, and estimating visibility, in both cases using static images of a scene over a period of time, only a portion of which are labeled.

2. REGULARIZATION WITH TEMPORAL PRIORS

Assume we have a training set of N labeled points $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^L$ and $\mathbf{y}_n \in \mathbb{R}^D$, and M additional unlabeled points $\{\mathbf{x}_m\}_{m=1}^M$. All $M+N$ inputs come as a collection of sequences of the form $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots)$, each of which is only partially labeled with $\mathbf{y}$-values. We want to estimate a regression mapping $\mathbf{f}$ that predicts the label $\mathbf{y}$ for an input $\mathbf{x}$ .We consider a least-squares regression setting with a graph Laplacian regularization [1, 2]:

$$E(\mathbf{f}) = \sum_{n=1}^N \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2 + \gamma_A \|\mathbf{f}\|_K^2 + \gamma_I \|\mathbf{f}\|_G^2 \quad (1)$$

where the $\|\mathbf{f}\|_K^2$ term refers to an RKHS norm (which encourages smoothness irrespectively of the training data distribution), and the $\|\mathbf{f}\|_G^2$ term refers to the graph Laplacian (which encourages smoothness of $\mathbf{f}$ with respect to the distribution of both labeled and unlabeled training points). The graph Laplacian is $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{W}$ is a given affinity matrix of $(N+M)\times(N+M)$ (such as the adjacency matrix or a Gaussian affinity matrix), and $\mathbf{D} = \operatorname{diag}(\mathbf{W}\mathbf{1})$ is the degree matrix (where $\mathbf{1}$ is a column vector of ones). Ordinarily, one might learn a neighborhood graph in an unsupervised way, such as the k-nearest-neighbor graph for a suitable value of k, but for the reasons mentioned in the introduction, this might be a poor regularizer (which encourages nearby points $\mathbf{x}$ to have similar $\mathbf{y}$-values, even though their true $\mathbf{y}$-values may be very different). Here we propose to construct the graph as the collection of input sequences in $\mathbf{x}$-space, *under the assumption that* $\mathbf{y}$ *varies slowly as a function of time but not necessarily as a function of* $\mathbf{x}$. Thus, if all $M+N$ points are indexed in order by sequence, the corresponding adjacency matrix will contain ones in the sub- and super-diagonals (except at a sequence end or start), and zeroes elsewhere. The sequential regularization term is then quadratic on the label values $\mathbf{f}(\mathbf{x}_n)$:

$$\|\mathbf{f}\|_{G_t}^2 = \mathbf{f}^T \mathbf{L}_t \mathbf{f} = \sum_{\substack{\text{sequen-}\\ \text{ces } s}} \sum_{n=2}^{N_s} w_{n,n-1}^s \|\mathbf{f}(\mathbf{x}_n^s) - \mathbf{f}(\mathbf{x}_{n-1}^s)\|^2 \quad (2)$$

which is to be compared with the usual graph Laplacian term:

$$\|\mathbf{f}\|_G^2 = \mathbf{f}^T \mathbf{L} \mathbf{f} = \sum_{n\sim m} w_{nm} \|\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{x}_m)\|^2. \quad (3)$$

Fig. 2: What are the angles of these digits? Are they sixes at x degrees or nines at $x + 180$ degrees?

Fig. 3: The 30 sixes and 30 nines from the MNIST dataset used to derive the rotated sequences.

Note this is not an additional graph prior term to the k-nearest-neighbor Laplacian, but that it replaces it. Although it is possible to use higher-order temporal priors (i.e., more neighbors within each sequence), in this paper we focus on linking consecutive points only. Our regularization term has the obvious interpretation of the squared gradient of $\mathbf{f}$ integrated along each sequence; thus, it cares about directional derivatives along paths in $\mathbb{R}^L$ rather than about the full gradient (which may be poorly constructed from the available samples in the problems we consider in this paper).

If the points along a sequence were also nearest neighbors, then the k-nearest-neighbor graph with a low value of k (perhaps even $k = 1$) would coincide or be very similar to our sequential graphs. However, in the problems we consider, the sequential structure is obscured by other neighboring points that greatly differ in label value. This is clear in our experiments, where the best value of k with the Laplacian prior is not very small. That is, for this type of problems, no value of k (or ϵ if we use an ϵ-ball graph) will yield a graph similar to the sequential graph. (In our experiments we have focused on 1D output values, but method carries over to multidimensional outputs in a straighforward way.)

The solution of this regularized least squares problem is analogous to the original one but replacing the graph Laplacian matrix with $\mathbf{L}_t$, the temporal graph Laplacian constructed from our sequences. The solution is unique and is given by a basis function expansion (depending on the RKHS) at each of the labeled and unlabeled points, $\mathbf{f}(\mathbf{x}) = \sum_{n=1}^{N+M} \boldsymbol{\alpha}_n K(\mathbf{x}_n, \mathbf{x})$, and the weights $\boldsymbol{\alpha}_n$ of this expansion are given by the solution of a linear system of size $(M+N) \times (M+N)$. In this paper we use Gaussian kernels $K(\cdot, \cdot)$ of width σ. As described, we fix the loss function to the squared regression error, which results in a simple algorithm. Other formulations are possible with our regularization, such as a hinge loss.

3. EXPERIMENTS

We performed three experiments using both synthetic and real world image datasets. Each experiment compares the results of the proposed method, temporal Laplacian regularized least squares regression (TLapRLSR) of eq. (2), with standard Laplacian regularized least squares regression (LapRLSR) of eq. (3). Gaussian radial basis function (RBF) kernels are used in all cases.

3.1. Image data: rotated MNIST digits

This experiment investigates the challenging task of estimating the orientations of handwritten digits that are very similar except for a fixed rotation. For example, consider the digits in figure 2. It is difficult even for a human observer to tell whether the digits are a six at x degrees or a nine at $x + 180$ degrees. Our dataset consists of rotated versions of 30 sixes and 30 nines from the MNIST database shown in figure 3. Each of the 30 nines are rotated counter-clockwise at one degree intervals from 0 to 180 degrees (labels are 0 to 180 degrees) and each of the 30 sixes are rotated counter-clockwise at one degree intervals from 180 to 360 degrees (labels are 180 to 360 degrees) for a total of 10 860 images. This results in a dataset in which *a six might appear very similar to a nine except for a 180 degree phase difference*. Therefore, Euclidean distances in image space would consider an upright 9 and an upside-down 6 as neighbors even though their

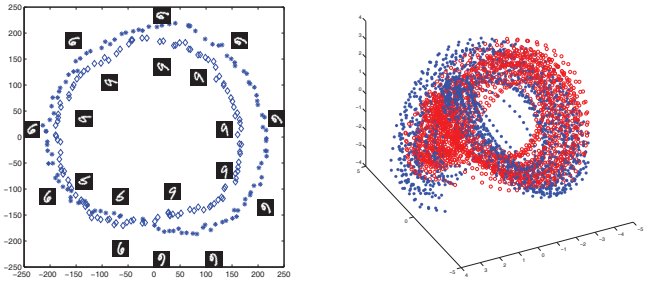


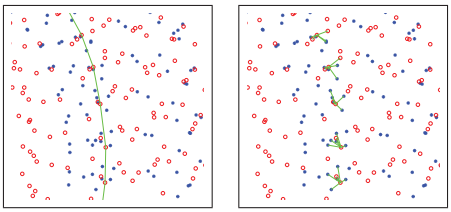
Fig. 4: *Left*: images corresponding to a rotated six (asterisks $*$) and a rotated nine (diamonds $\blacklozenge$) sequence projected onto the first two PCA components. Both sequences start at the bottom and go counter-clockwise. The sequences are actually (parallel) spirals in the full three dimensional PCA space. *Right*: the rotated sequences of sixes (red) and nines (blue) projected onto the first three PCA components.

labels differ drastically. Figure 4a demonstrates this for a particular pair of six and nine sequences.

The every-degree dataset is subsampled at three and five degrees to create two additional datasets for investigating the effect of the relation of the within-to-between sequence distances. Training, cross-validation and evaluation sets are constructed from the 60 sequences as follows. The elements of each sequence are assigned in an alternating fashion to training, cross-validation and evaluation sets resulting in a training set containing $61 \times 60 = 3\,660$ images, and cross-validation and evaluation sets containing $60 \times 60 = 3\,600$ images each for the every-degree dataset. This is reduced to 1 220–1 200–1 200 and 732–700–700 for the every three and every five degree datasets, respectively. A percentage of the training set is labeled (with the rotation angle) at approximately equal spacings with respect to angle. The 60 sequences of labeled and unlabeled images in the training set form the "temporal" sequences for TLapRLSR.

Comparisons are performed using both the native feature space, in which the distance between images is the square-root of the sum of the squares of the pixel differences (Euclidean distance between the images treated as vectors), as well as in a reduced 3D space, in which the images are projected onto the first three principal components computed over the training set and the difference between two images is the square-root of the sum of squares of the projection differences (Euclidean distance between the projected values). We refer to this reduced space as the principal component analysis (PCA) space. Figure 4a shows the distribution of the images corresponding to a particular pair of six and nine sequences in the two dimensional space formed by the first two principal components. Note again how samples from the different sequences can be closer to each other in the feature space than they are to samples from the same sequence even though there is a 180 degree phase shift. This scenario becomes even more likely with the full 60 sequences. Figure 4b plots all 60 sequences in the 3D PCA space. Note how the sixes' and nines' sequences interleave in a complex way, making it extremely hard to estimate the rotation angle. Although the angle along a sequence does vary smoothly, slight deviations outside the sequence can produce large angle deviations, and the smoothness assumptions built into a usual graph Laplacian may not hold well here. Figure 5 shows a 2D view of how the different approaches construct the graphs: the k-nearest-neighbor graph links both sixes and nines (thus encouraging them to have the same label, even though it differs by 180 degrees), while the temporal graph does not make that mistake.

We determined optimal parameter values using cross-validation. For both the temporal and k-nn approaches we set $\gamma_A = 10^{-3}$ (this value was shown to give reasonable results for both approaches). For



Circles ○: sixes; asterisks ∗: nines; edges in green

Fig. 5: Examples of graphs constructed for the rotated digits problem by the proposed approach using the sequence information (left) and the standard approach with 6 nearest neighbors (right).

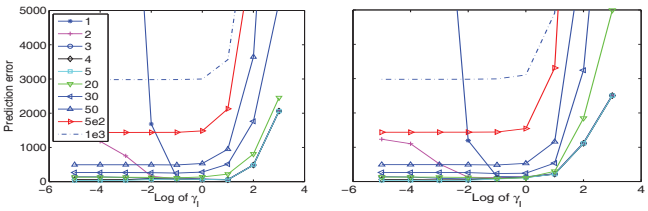


Fig. 6: Prediction errors in regions bracketing the optimal free parameter values for TLapRLSR (left) and LapRLSR (right) for the every-one-degree rotated digit problem in the native space.

the k-nn approach we initially set $k = 6$ for constructing the adjacency matrix following the approach of [1]. Optimal values for the remaining two parameters, γ_I and σ, the width of the RBF kernel, are determined using the training and cross-validation sets. Figure 6 shows the prediction error for ranges of values of these parameters for the temporal and k-nn approaches for the native space for the rotated every one degree dataset. The plots for the PCA space are similar. It is clear that the temporal information reduces the prediction error, especially towards the right of the plots, where larger values of γ_I mean the graph Laplacian is more heavily weighed.

Figure 7 compares the predicted versus true angles for the test images for the every-one-degree rotated digits. It is very remarkable that TLapRLSR yields almost perfect prediction except in two small areas where it makes some large errors, while LapRLSR makes quite large errors almost everywhere. This results in TLapRLSR having always a smaller mean absolute error, even though sometimes LapRLSR does achieve the better mean squared error. Figure 10 lists the prediction errors achieved by the optimal parameter settings and for different numbers of neighbors k for the LapRLSR approach. These results correspond to a training set in which 20% of the images are labeled. Figure 8 shows the dependence of the two approaches on the percentage of labeled data for the every one degree rotated digits problem. Also shown is the prediction error for the fully supervised case, in which only the labeled training data is used (this is equivalent to setting $\gamma_I = 0$, effectively dropping the graph Laplacian term from the objective function). Learning is performed ten times for each ratio value using randomly perturbed training sets. The error bars in the plot indicate the standard deviation of the prediction error. Note that our proposed TLapRLSR approach results in a significantly lower prediction error for a broad range of labeled data ratios. The margin of improvement often increases as the ratio of labeled data decreases which makes the proposed approach particularly attractive for the practically important problems were there is very little labeled data. The increase for TLapRLSR for high ratios for the PCA space indicates that the value γ_I that was found to be optimal for a ratio of 0.2 starts to penalize the proposed approach as the ratio of labeled points increases.

3.2. Image data: estimating time of day

This experiment investigates the problem of estimating the time of day at which an image is captured. The dataset consists of 870

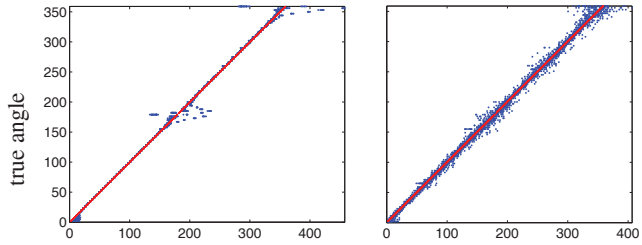


Fig. 7: True (Y–axis) vs predicted (X) angle on test images for TLapRLSR (left) and LapRLSR (right) for the every-one-degree rotated digits problem. The nines are plotted from 0° to 180° and the sixes from 180° to 360°. Perfect performance: diagonal red line.

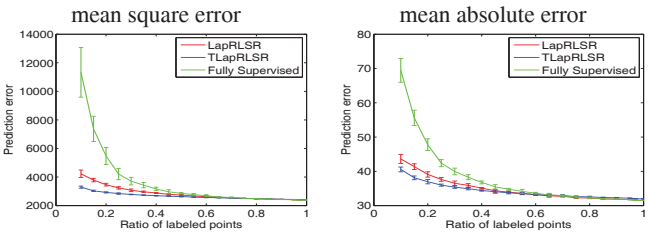


Fig. 8: Dependence of the prediction error on the ratio of labeled points in the every-one-degree rotated digit problem (PCA space).

grayscale images of an outdoor scene acquired by a static camera at one minute intervals from 4:40am to 7:15pm. Figure 1 shows images from 5:49 am, 5:20 pm, and 5:49 pm. The images are approximately evenly divided into training, cross-validation, and evaluation datasets. Approximately 20% of the training images are labeled (with the time of day) at equal spaced time intervals. The images in the training set form the single temporal sequence for TLapRLSR. Temporal information is critical for incorporating unlabeled images into the learning process for this problem since (1) parts of the scene might change rapidly over short time intervals due to clouds that obscure the sun or enter the scene, other objects, etc., and (2) images spaced far apart in time, such as from the morning and afternoon, can look similar.

Again, comparisons are performed using both the native and reduced image spaces. The images are resized to 64×64 pixels. We set $\gamma_A = 10^{-3}$ and $k = 6$, and determine optimal values for the remaining two parameters, γ_I and σ, through cross-validation. Figure 10 lists the minimum prediction errors achieved by the optimal parameter settings. TLapRLSR is again shown to outperform LapRLSR, both in mean squared error and mean absolute error, by a large margin (error two or three times smaller).

3.3. Image data: estimating visibility

We investigate the problem of estimating atmospheric visibility from images of scenes with objects at a range of distances. Atmospheric visibility is typically measured using a transmissometer which computes the extinction coefficient b_{ext} of the atmosphere based on the attenuation of a laser beam transmitted from an emitter to a receiver spaced kilometers apart. b_{ext} is defined as the fractional attenuation of light per unit distance and is reported in terms of inverse distance such as inverse megameters (Mm^{-1}). Transmissometers are expensive instruments that require accurate calibration and so the option of using commodity digital cameras is appealing even if the measurements are not as accurate. The challenge is to learn the highly nonlinear mapping from the complex image space to the b_{ext} values.

Our dataset consists of grayscale images of the Phoenix, Arizona region taken every 15 minutes between 8am and 5pm for two weeks. The images are captured at 0, 15, 30, and 45 minutes past

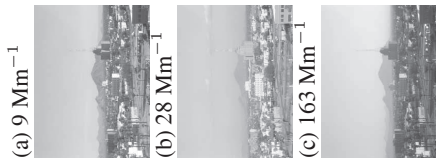


Fig. 9: Images showing different visibility levels ((a) good, (b) moderate, (c) poor) according to the coefficient of extinction b_{ext} as measured with a transmissometer.

	Rotated Digits						Time of Day		Visibility	
	Every 1 Degree		Every 3 Degrees		Every 5 Degrees					
	Native	PCA	Native	PCA	Native	PCA	Native	PCA	Native	PCA
TLap	92.7 (**2.59**)	**2703** (**35.6**)	301 (**10.3**)	**3670** (**47.9**)	529 (**15.3**)	**3880** (**45.9**)	**0.111**	**0.568**	**116.7**	**159.2**
Lap k=1	189 (8.47)	3330 (40.1)	331 (13.0)	4360 (50.9)	2140 (33.3)	4650 (53.5)	0.346	1.07	120.8	161.6
Lap k=2	32.9 (2.88)	3180 (38.5)	**240** (10.9)	4150 (49.4)	1082 (23.5)	4590 (53.2)	0.330	1.02	119.0	160.0
Lap k=3	49.2 (4.26)	3130 (38.0)	258 (11.3)	4080 (49.1)	716 (10.5)	4540 (53.1)	0.329	1.03	119.5	159.8
Lap k=4	**27.3** (2.87)	3090 (37.5)	279 (11.6)	4060 (49.0)	613 (18.6)	4500 (53.0)	0.332	0.967	120.4	159.9
Lap k=5	42.7 (3.92)	3080 (37.4)	282 (11.7)	4040 (49.0)	541 (17.7)	4500 (53.0)	0.345	0.930	119.6	159.9
Lap k=6	41.7 (3.82)	3060 (37.2)	295 (12.1)	4030 (49.0)	494 (16.9)	4500 (53.0)	0.348	0.930	118.3	160.1
Lap k=7	52.0 (4.42)	3040 (37.1)	294 (12.1)	4020 (49.0)	**449** (15.9)	4480 (53.0)	0.349	0.995	118.0	160.4

Fig. 10: Prediction errors with optimal parameter settings: mean square (mean absolute).

the hour. We also have transmissometer measurements of b_{ext} over the same period but only at the top of each hour. Figure 9 shows images corresponding to different values of b_{ext}. Note the complexity of the scene and that there are potentially lots of factors that can cause changes in the scene besides atmospheric scattering and absorption. Or, equivalently, two images might look similar in some regions but have different b_{ext} values. The temporal regularization method proposed in this paper is motivated by the problem of how to include the unlabeled images, those for which we do not have concurrent transmissometer readings, in the learning process.

There are 457 images in our dataset (some images are missing) of which 123 are labeled. We set aside every other labeled image as the test dataset. The remaining labeled and unlabeled images form the 14 temporal sequences for TLapRLSR. We set $\gamma_A = 10^{-3}$ and $k = 6$, and determine optimal values for the remaining two parameters, γ_I and σ, through cross validation on the test set. Figure 10 lists the minimum prediction errors achieved by the optimal parameter settings. The proposed approach again outperforms the standard approach although this time by less of a margin.

4. RELATED WORK

Temporal priors have been used in many models, however we focus on their application to semi-supervised learning and neighborhood graphs. Besides the output value or label of a pattern, different types of supervisory information have been proposed in the semi-supervised literature. Constraints (must-link, cannot-link) in clustering indicate more or less strongly whether a pair of patterns should be in the same or in different clusters (e.g. [3]). Rank order constraints [4] specify that the scalar output value of a pattern A should be greater or equal than that of a pattern B. This information may then be combined with a neighborhood graph that is learnt in an unsupervised way (k-nearest-neighbor or ϵ-ball) from the entire input patterns (labeled or unlabeled). In contrast, our supervisory information essentially consists of the neighborhood graph, in the form of disconnected sequences of input patterns (derived from a known temporal structure). While this information may not be useful in all situations or may not always be available, we have shown that it is very helpful when the unsupervised neighborhood information conflicts strongly with the output values or labels. In this case, using the unsupervised neighborhood information incorrectly regularizes the problem, and not using it makes the problem too unconstrained with sparse label information. Note that, as done in [4], order preferences can be used to encode similarities of the form $a \ll f(\mathbf{x}_n) - f(\mathbf{x}_m) \ll b$, and one could chain pairs of this form along a sequence. However, this is limited to scalar outputs, and the formulation using a sequential neighborhood graph is simpler.

Our work is also related to the tracking approach of [5]. They consider the problem of mapping a given time series, such as a video of a moving object, to a state space, such as the location of the object. Unlike our objective function, theirs depends also on all the unknown output values, which are taken as parameters, and a temporal prior is applied to them rather than to the predictor function. Our approach is simpler and has a more efficient, closed-form solution. Another related work is that of [6] for manifold learning (rather than semi-supervised regression). They apply Isomap to sequential data by modifying the k-nearest-neighbor graph to reduce the distance between temporally adjacent points. Our approach does not modify the k-nearest-neighbor graph but actually replaces it with the sequential one.

5. CONCLUSION

We have considered a semi-supervised regression setting where the supervision information consists not only of the labels at a small proportion of the input patterns, but of the neighborhood graph of the inputs as a collection of disconnected sequences. Our algorithm adapts Laplacian regularization for regression with this graph. We have shown it to improve consistently over using as regularizer the k-nearest-neighbor graph of the inputs in cases where this neighborhood information is not a good predictor of the label—for example, when nearby inputs may have very different labels. This is particularly useful for problems with temporal structure and high-dimensional complex inputs, such as images, where it is very hard to tell which of the many things that change from image to image have an effect on the label. We expect our approach would also be useful for classification problems of this type; and to problems where the neighborhood information is not temporal but of other types.

Acknowledgements. Carreira-Perpiñán funded by NSF award IIS-0754089 (CAREER). Newsam and Xie funded in part by the Center for Information Technology Research in the Interest of Society and a UC TSR&TP Atmospheric Aerosols and Health Graduate Fellowship. We would like to thank the Arizona Department of Environmental Quality and Air Resource Specialists, Inc., for providing the Phoenix image and transmissometer data.

6. REFERENCES

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7, 2006. 1, 3

[2] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. *AISTATS*, 2005. 1

[3] Z. Lu and M. Carreira-Perpiñán. Constrained spectral clustering through affinity propagation. *CVPR*, 2008. 4

[4] X. Zhu and A. Goldberg. Kernel regression with order preferences. *AAAI*, 2007. 4

[5] A. Rahimi, B. Recht, and T. Darrell. Learning appearance manifolds from video. *CVPR*, 2005. 4

[6] O. Jenkins and M. Matarić. A spatio-temporal extension to Isomap nonlinear dimension reduction. *ICML*, 2004. 4

[7] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *ICML*, 2003.

[8] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. *CVPR*, 2007. 1

[9] D. Baumer, S. Versick, and B. Vogel. Determination of the visibility using a digital panorama camera. *Atmospheric Environment*, 2008. 1

[10] F. M. Caimi, D.M. Kocak, and J. Justak. Remote visibility measurement technique using object plane data from digital image sensors. *IEEE Int. Geoscience & Remote Sensing Sym.*, 2004. 1