Kernel Methods for Weakly Supervised Mean Shift Clustering

Oncel Tuzel[(1)] Fatih Porikli[(1)] Peter Meer[(2)]
Mitsubishi Electric Research Laboratories[(1)]
Cambridge, MA 02139
Electrical and Computer Engineering[(2)]
Rutgers University, Piscataway, NJ 08854

Abstract

Mean shift clustering is a powerful unsupervised data analysis technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. The data association criteria is based on the underlying probability distribution of the data points which is defined in advance via the employed distance metric. In many problem domains, the initially designed distance metric fails to resolve the ambiguities in the clustering process. We present a novel semi-supervised kernel mean shift algorithm where the inherent structure of the data points is learned with a few user supplied constraints in addition to the original metric. The constraints we consider are the pairs of points that should be clustered together. The data points are implicitly mapped to a higher dimensional space induced by the kernel function where the constraints can be effectively enforced. The mode seeking is then performed on the embedded space and the approach preserves all the advantages of the original mean shift algorithm. Experiments on challenging synthetic and real data clearly demonstrate that significant improvements in clustering accuracy can be achieved by employing only a few constraints.

1. Introduction

Mean shift is an iterative procedure for locating the stationary points of a density function represented by a set of samples. Although the procedure was initially described decades ago [14], it's not been popular in vision community until its potential uses for feature space analysis and optimization were understood [7, 11]. Recently, the mean shift procedure is used for a wide range of computer vision applications such as visual tracking [8, 17], image smoothing and segmentation [10, 24], and information fusion [6, 9]. In addition, the theoretical properties of the procedure such as order and guarantee of convergence are discussed in several studies [4, 10, 13].

Mean shift clustering is an unsupervised density based nonparametric clustering technique. The data points are assumed to be originated from an unknown distribution which is approximated via kernel density estimation. The cluster centers are located by the mean shift procedure and the data points associated with the same local maxima of the density function (modes) produce a partitioning of the space.

In many cases, prior information about the problem domain is available in addition to the data instances. For example, a partial labeling of the data can be acquired from a secondary process (e.g. face detector for scene categorization) or via simple user supervision (e.g. a human operator for initial segmentation of tumors).

Recently, semi-supervised approaches that aim to incorporate prior information and labeled data into the clustering algorithm as a guide have received considerable attention in machine learning and computer vision [1, 5, 15] including the background constrained k-means [23], adaptive kernel k-means [25], and kernel graph clustering [18]. It is shown that unlabeled data, when used in conjunction with a small amount of labeled data, can produce significant improvement in clustering accuracy. Most semi-supervised clustering algorithms assume that pairwise must-link constraints, i.e. pairs of points that should belong in the same cluster, are provided with the data. Transitive binary constraints of this form are natural in the context of the graph partitioning where edges in the graph encode pairwise relationships as in graph cuts [2] and random walk segmentation [16].

Unlike semi-supervised variants of k-means, spectral, and graph clustering methods, the existing mean shift methods do not have a mechanism to utilize the labeled prior information in order to guide the type of the clusters desired. Such a mean shift method would be critical in many applications from scene classification with weakly labeled data to image segmentation.

In this paper we present a semi-supervised mean shift clustering. To our knowledge, this is the first method that incorporates the prior information into the mean shift clustering by implicitly mapping the must-link constraints and objective function into a higher dimensional space induced by a kernel function as described in the following sections.

2. Method Overview

The motivation in this paper is to enforce a set of constraints given in terms of the pairwise similarities such that the result of the final clustering algorithm groups the con-

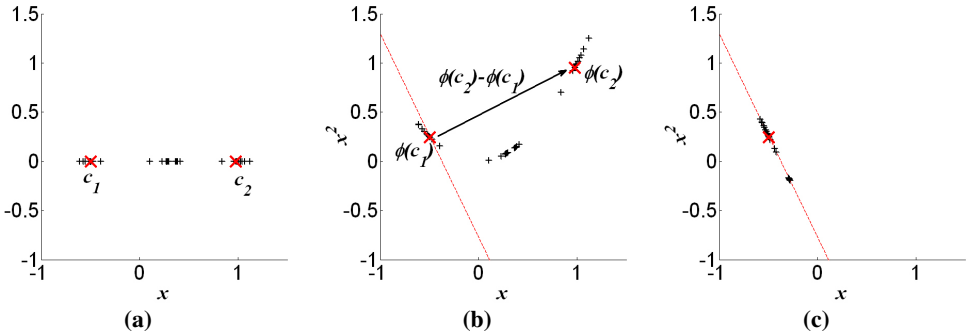

Figure 1. Illustration of constrained clustering. **(a)** Input space. Red crosses mark the constraint pair $(\mathbf{c}_1, \mathbf{c}_2)$. **(b)** The input space is mapped to the feature space via quadratic mapping $\phi(x) = [x \;\; x^2]^T$. The black arrow is the constraint vector $(\phi(\mathbf{c}_2) - \phi(\mathbf{c}_1))^T$, and the red line is its null space. **(c)** The feature space is projected to the null space of the constraint vector. Constraint points collapse to a single point therefore the clustering algorithm trivially groups them together. Two clusters can be easily identified.

strained point pairs into the same clusters. In addition, we would like to not only modify the clusters at the local scale (just for the constraint pairs) but carry the enforced structure to the entire input space. By doing so, it is possible to guide the clustering towards the interested structure of the input space using only a few constraint.

The proposed approach for constrained mean shift clustering is based on embedding the input space to a space where the constraint pairs are associated with the same mode when the density estimation is performed on the embedded space. In addition we would like to enforce that the original distances are best preserved while satisfying constraints. The proposed method still preserves all the advantages of the original mean shift clustering since it is equivalent to applying the procedure on the embedded space.

We start with the base example where a single constraint is given and the clusters are linearly separable. Let $(\mathbf{c}_1, \mathbf{c}_2)$ be the point pair that is constrained to be clustered together. If we project the input space to the null space of $(\mathbf{c}_2 - \mathbf{c}_1)^T$ which are the orthogonal directions to the constraint vector, the points $\mathbf{c}_1$ and $\mathbf{c}_2$ map to the same point, therefore it is guaranteed that they are associated with the same mode. In addition, null space projection is the optimal linear projection in the sense that it preserves the variance along the orthogonal directions to the projection direction, hence the original distance measure is best preserved.

This approach does not scale well with the increasing number of constraints. Notice that, given m linearly independent constraint vectors on a d-dimensional input space, the null space of the constraint matrix is $d-m$ dimensional. This implies that if more than $d-1$ constraints are specified all the points collapse to a single point therefore clustered together. A simple solution exists for this problem and for the generalization to the linearly inseparable case using a mapping function ϕ which embeds the input space to an enlarged feature space. On the embedded space the same technique can be applied by projecting the points to the null space of $(\phi(\mathbf{c}_2) - \phi(\mathbf{c}_1))^T$.

In Figure 1, we present a simple illustration for a one-dimensional example. Data on the input space appears to be originated from three clusters. We incorporate the prior information in terms of the pairwise constraint enforcing the two cross marked points to be clustered together (Figure 1a). In Figure 1b, data is explicitly mapped to two-dimensional space via the quadratic mapping $\phi(x) = [x \;\; x^2]^T$. This mapping is arbitrary and only used for illustration purpose. The black arrow denotes the constraint vector and the red dashed line denotes its null space. By projecting the input space to the null space of the constraint vector the constraint points collapse to a single point therefore the clustering algorithm trivially groups them together (Figure 1c). In addition, the projection carries the enforced structure to the entire feature space and there are two significant clusters.

Explicitly designing the mapping function and working on high dimensional spaces is not practical. We present a kernel based mean shift algorithm which implicitly works on the enlarged feature space and extends the Euclidean mean shift to inner product spaces. The constrained mean shift on the kernel induced space then reduces to modifying the kernel matrix with respect to the defined constraints.

3. Mean Shift Clustering

In this section we briefly describe the variable bandwidth mean shift procedure [9]. Given n data points $\mathbf{x}_i$ on a d-dimensional space $\mathbb{R}^d$ and the associated bandwidths $h_i = h(\mathbf{x}_i)$, $i = 1, ..., n$, the sample point density estimator obtained with profile $k(\mathbf{x})$ is given by

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^d} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h_i} \right\|^2 \right). \tag{1}$$

We utilize multivariate normal profile

$$k(x) = e^{-\frac{1}{2}x} \qquad x \geq 0. \tag{2}$$

Taking the gradient of (1), we observe that the stationary points of the density function satisfy

$$\frac{2}{n}\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}(\mathbf{x}_i-\mathbf{x})g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right)=0 \qquad (3)$$

where $g(x)=-k'(x)$. The solution can be found iteratively via the fixed point algorithm

$$\bar{\mathbf{x}}=\frac{\sum_{i=1}^{n}\frac{\mathbf{x}_i}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right)} \qquad (4)$$

which is called mean shift procedure. Comaniciu and Meer [10] show that the convergence to a local mode of the distribution is guaranteed when the mean shift iterations are started at a data point.

4. Kernel Mean Shift

In this section we present the kernel mean shift algorithm which forms the basis for the constrained mean shift algorithm. The presented approach also extends the original mean shift algorithm from Euclidean spaces to inner product spaces thereby it is possible to apply the algorithm to larger class of problem domains such as clustering on manifolds [21]. However we will not focus on this aspect since it is beyond the scope of the paper. We also note that a similar derivation for fixed bandwidth mean shift algorithm was given in [22].

Let $\mathcal{X}$ be the input space such that, $\mathbf{x}_i\in\mathcal{X}$, $i=1,...,n$. It is not necessary that the input space is an Euclidean space $\mathbb{R}^d$. Let $K:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ be a positive definite kernel function satisfying for all $\mathbf{x},\mathbf{x}'\in\mathcal{X}$

$$K(\mathbf{x},\mathbf{x}')=\boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \qquad (5)$$

where $\boldsymbol{\phi}$ maps input space into the $d_{\boldsymbol{\phi}}$-dimensional feature space $\mathcal{H}$, $\boldsymbol{\phi}(\mathbf{x})=[\phi_1(\mathbf{x})\ \ \phi_2(\mathbf{x})\ \ \dots\ \ \phi_{d_\phi}(\mathbf{x})]^T$. The use of kernel makes it possible to map the data implicitly to an enlarged feature space where the nonlinear structure of the data points can be studied and the constraints can be effectively applied.

We first derive the mean shift procedure on the feature space $\mathcal{H}$ in terms of the explicit representation of the mapping $\boldsymbol{\phi}$. The point sample density estimator at $\mathbf{y}\in\mathcal{H}$ is given by

$$f_{\mathcal{H}}(\mathbf{y})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^{d_\phi}}k\left(\left\|\frac{\mathbf{y}-\boldsymbol{\phi}(\mathbf{x}_i)}{h_i}\right\|^2\right) \qquad (6)$$

Taking the gradient of (6) with respect to $\boldsymbol{\phi}$, the stationary points of the density function satisfy

$$\frac{2}{n}\sum_{i=1}^{n}\frac{1}{h_i^{d_{\boldsymbol{\phi}}+2}}(\boldsymbol{\phi}(\mathbf{x}_i)-\mathbf{y})g\left(\left\|\frac{\mathbf{y}-\boldsymbol{\phi}(\mathbf{x}_i)}{h_i}\right\|^2\right)=0. \qquad (7)$$

Like (4) the solution can be found iteratively

$$\bar{\mathbf{y}}=\frac{\sum_{i=1}^{n}\frac{\boldsymbol{\phi}(\mathbf{x}_i)}{h_i^{d_{\boldsymbol{\phi}}+2}}g\left(\left\|\frac{\mathbf{y}-\boldsymbol{\phi}(\mathbf{x}_i)}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d_{\boldsymbol{\phi}}+2}}g\left(\left\|\frac{\mathbf{y}-\boldsymbol{\phi}(\mathbf{x}_i)}{h_i}\right\|^2\right)}. \qquad (8)$$

Now we derive the implicit form of the algorithm. Let

$$\boldsymbol{\Phi}=[\boldsymbol{\phi}(\mathbf{x}_1)\ \ \boldsymbol{\phi}(\mathbf{x}_2)\ \ \dots\ \ \boldsymbol{\phi}(\mathbf{x}_n)] \qquad (9)$$

be the $d_{\boldsymbol{\phi}}\times n$ matrix of the feature points and $\mathbf{K}=\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ be the $n\times n$ Kernel (Gram) matrix. We observe that at each iteration of the mean shift procedure (8), the estimate $\bar{\mathbf{y}}$ lies in the column space of $\boldsymbol{\Phi}$. Any point on the subspace spanned by the columns of $\boldsymbol{\Phi}$ can be written as

$$\mathbf{y}=\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}} \qquad (10)$$

where $\boldsymbol{\alpha}_{\mathbf{y}}$ is an n-dimensional weighting vector. The distance between two points $\mathbf{y}$ and $\mathbf{y}'$ on the subspace is

$$\begin{aligned}\|\mathbf{y}-\mathbf{y}'\|^2 &= \|\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}}-\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}'}\|^2 \qquad (11)\\ &= \boldsymbol{\alpha}_{\mathbf{y}}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}}+\boldsymbol{\alpha}_{\mathbf{y}'}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}'}-2\boldsymbol{\alpha}_{\mathbf{y}}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}'}\\ &= \boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\boldsymbol{\alpha}_{\mathbf{y}}+\boldsymbol{\alpha}_{\mathbf{y}'}^T\mathbf{K}\boldsymbol{\alpha}_{\mathbf{y}'}-2\boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\boldsymbol{\alpha}_{\mathbf{y}'}.\end{aligned}$$

The distances can be expressed in terms of the inner product of the points and the algorithm iteratively updates the weighting vector $\boldsymbol{\alpha}_{\mathbf{y}}$. Let $\mathbf{e}_i$ denote the i-th canonical basis for $\mathbb{R}^n$. Substituting (11) into (8), and using the equivalence $\boldsymbol{\phi}(\mathbf{x}_i)=\boldsymbol{\Phi}\mathbf{e}_i$, the mean shift procedure becomes

$$\bar{\boldsymbol{\alpha}}_{\mathbf{y}}=\frac{\sum_{i=1}^{n}\frac{\mathbf{e}_i}{h_i^{d_{\boldsymbol{\phi}}+2}}g\left(\frac{\boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\boldsymbol{\alpha}_{\mathbf{y}}+\mathbf{e}_i^T\mathbf{K}\mathbf{e}_i-2\boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\mathbf{e}_i}{h_i^2}\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d_{\boldsymbol{\phi}}+2}}g\left(\frac{\boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\boldsymbol{\alpha}_{\mathbf{y}}+\mathbf{e}_i^T\mathbf{K}\mathbf{e}_i-2\boldsymbol{\alpha}_{\mathbf{y}}^T\mathbf{K}\mathbf{e}_i}{h_i^2}\right)}. \qquad (12)$$

The clustering algorithm starts on the data points on $\mathcal{H}$, therefore the initial weighting vectors are given by $\boldsymbol{\alpha}_{\mathbf{y}_i}=\mathbf{e}_i$, such that, $\mathbf{y}_i=\boldsymbol{\Phi}\boldsymbol{\alpha}_{\mathbf{y}_i}=\boldsymbol{\phi}(\mathbf{x}_i)$, $i=1\dots n$. Upon convergence the mode can be expressed via $\boldsymbol{\Phi}\bar{\boldsymbol{\alpha}}_{\mathbf{y}_i}$. The points converging to the same mode are clustered together. The convergence of the procedure follows from the original proof [10], since any positive semidefinite matrix $\mathbf{K}$ is a kernel for some feature space [12] and the derived method implicitly applies mean shift on that feature space.

Note that, when the rank of the kernel matrix $\mathbf{K}$ is smaller than n, columns of $\boldsymbol{\Phi}$ form an overcomplete basis. Therefore the modes can be identified within an equivalence relationship where two modes $\boldsymbol{\Phi}\bar{\boldsymbol{\alpha}}_{\mathbf{y}}$ and $\boldsymbol{\Phi}\bar{\boldsymbol{\alpha}}_{\mathbf{y}'}$ are identified as same when $\|\boldsymbol{\Phi}\bar{\boldsymbol{\alpha}}_{\mathbf{y}}-\boldsymbol{\Phi}\bar{\boldsymbol{\alpha}}_{\mathbf{y}'}\|^2=\bar{\boldsymbol{\alpha}}_{\mathbf{y}}^T\mathbf{K}\bar{\boldsymbol{\alpha}}_{\mathbf{y}}+\bar{\boldsymbol{\alpha}}_{\mathbf{y}'}^T\mathbf{K}\bar{\boldsymbol{\alpha}}_{\mathbf{y}'}-2\bar{\boldsymbol{\alpha}}_{\mathbf{y}}^T\mathbf{K}\bar{\boldsymbol{\alpha}}_{\mathbf{y}'}=0$. In addition, it is possible that the mode can not be represented on the input space $\mathcal{X}$ since the mapping $\boldsymbol{\phi}$ is not necessarily invertible. Still it is possible to compute the distance of any point on the input space to the modes using (11).

Another important point is that the dimensionality of the feature space $d_{\boldsymbol{\phi}}$ can be very large. For instance, it is infinite dimensional, if the original kernel function (5) is the Gaussian kernel. In those cases it may not be possible to compute the point sample density estimator (6), consequently the mean shift procedure (if h_i is not constant). However, the procedure is restricted to the subspace spanned by the feature points. Since we have limited number of samples n, the dimensionality of the subspace can be estimated from the rank of the kernel matrix $\mathbf{K}$.

5. Constrained Mean Shift Clustering

Let $\{(\mathbf{c}_{j,1}, \mathbf{c}_{j,2})\}_{j=1\dots m}$ be the set of m input point pairs which are identified to be clustered together. Following the discussion presented in Section 2, we initially embed the input space to a feature space via mapping $\boldsymbol{\phi}$ and the constraints are satisfied by projecting the feature space to the null space of the constraint vectors.

Let $\mathbf{A}$ be the $m \times d_{\boldsymbol{\phi}}$ dimensional constraint matrix

$$\mathbf{A} = \begin{pmatrix} (\boldsymbol{\phi}(\mathbf{c}_{1,1}) - \boldsymbol{\phi}(\mathbf{c}_{1,2}))^T \\ \vdots \\ (\boldsymbol{\phi}(\mathbf{c}_{m,1}) - \boldsymbol{\phi}(\mathbf{c}_{m,2}))^T \end{pmatrix}. \tag{13}$$

The null space of $\mathbf{A}$ is the set of vectors $N[\mathbf{A}] = \{\mathbf{w} \in \mathcal{H} : \mathbf{A}\mathbf{w} = 0\}$. The matrix

$$\mathbf{P} = \mathbf{I}_{d_{\boldsymbol{\phi}}} - \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^+ \mathbf{A} \tag{14}$$

projects the vectors in $\mathcal{H}$ to the $N[\mathbf{A}]$ where $\mathbf{I}_{d_{\boldsymbol{\phi}}}$ is the $d_{\boldsymbol{\phi}}$ dimensional identity matrix and $^+$ is the pseudo inverse operation. The null space projection matrix satisfies $\mathbf{P}\boldsymbol{\phi}(\mathbf{c}_{j,1}) = \mathbf{P}\boldsymbol{\phi}(\mathbf{c}_{j,2}), j = 1 \dots m$.

Let $\mathbf{S} = \mathbf{A}\mathbf{A}^T$ be the $m \times m$ scaling matrix. The entries of $\mathbf{S}$ only involves the feature points through the inner product (5)

$$\begin{aligned} \mathbf{S}_{i,j} &= K(\mathbf{c}_{i,1}, \mathbf{c}_{j,1}) - K(\mathbf{c}_{i,1}, \mathbf{c}_{j,2}) \\ &\quad -K(\mathbf{c}_{i,2}, \mathbf{c}_{j,1}) + K(\mathbf{c}_{i,2}, \mathbf{c}_{j,2}). \end{aligned} \tag{15}$$

Therefore $\mathbf{S}^+$ can be computed without knowing the mapping $\boldsymbol{\phi}$.

Given the data points and the constraint set, the constrained mean shift algorithm maps the data points to the null space of the constraint matrix

$$\hat{\boldsymbol{\phi}}(\mathbf{x}) = \mathbf{P}\boldsymbol{\phi}(\mathbf{x}) \tag{16}$$

and implicitly performs mean shift on the embedded space. Since the constraint point pairs map to the same feature point, it is guaranteed that they converge to the same mode.

Instead of rewriting the mean shift procedure on the embedded space, it suffices to modify the kernel matrix $\mathbf{K}$ with respect to the projection and apply the derived kernel mean shift algorithm (12) on the modified kernel matrix $\hat{\mathbf{K}}$. The equivalence simply follows from the fact that apart from the distance computations the procedure is identical and the distances only involve feature points in terms of the inner products.

The projected kernel function is given by

$$\begin{aligned} \hat{K}(\mathbf{x}, \mathbf{x}') &= \hat{\boldsymbol{\phi}}(\mathbf{x})^T \hat{\boldsymbol{\phi}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T \mathbf{P}^T \mathbf{P} \boldsymbol{\phi}(\mathbf{x}') \\ &= \boldsymbol{\phi}(\mathbf{x})^T \mathbf{P} \boldsymbol{\phi}(\mathbf{x}') \\ &= \boldsymbol{\phi}(\mathbf{x})^T (\mathbf{I}_{d_{\boldsymbol{\phi}}} - \mathbf{A}^T \, \mathbf{S}^+ \mathbf{A}) \boldsymbol{\phi}(\mathbf{x}') \\ &= K(\mathbf{x}, \mathbf{x}') - K(\boldsymbol{\phi}(\mathbf{x}), \mathbf{A})^T \mathbf{S}^+ K(\boldsymbol{\phi}(\mathbf{x}'), \mathbf{A}). \end{aligned} \tag{17}$$

The identity $\mathbf{P}^T\mathbf{P} = \mathbf{P}$ follows from the fact that $\mathbf{P}$ is a projection matrix and it is symmetric. With a slight abuse of notation $K(\boldsymbol{\phi}(\mathbf{x}), \mathbf{A})$ denotes the m-dimensional vector

$$\begin{pmatrix} K(\mathbf{x}, \mathbf{c}_{1,1}) - K(\mathbf{x}, \mathbf{c}_{1,2}) \\ \vdots \\ K(\mathbf{x}, \mathbf{c}_{m,1}) - K(\mathbf{x}, \mathbf{c}_{m,2}) \end{pmatrix}. \tag{18}$$

We see that the modified kernel function $\hat{K}(\mathbf{x}, \mathbf{x}')$ involves the mapping $\boldsymbol{\phi}$ via the inner products and can be written in terms of the original kernel function $K(\mathbf{x}, \mathbf{x}')$. Notice that the dimensionality of the projected subspace is m smaller than the original subspace if the constraints are linearly independent or equivalently $\mathbf{S}$ is full rank.

It is possible that, instead of the data points, user can only supply the kernel matrix $\mathbf{K}$ and the indices of the points for the constraints. Let $\boldsymbol{\nu}$ be the $m \times 2$ indexing matrix mapping the constraint set to the input points, such that $\mathbf{c}_{j,1} = \mathbf{x}_{\nu_{j,1}}$ and $\mathbf{c}_{j,2} = \mathbf{x}_{\nu_{j,2}}$, $j = 1 \dots m$. We refer to the i,j-th element of a matrix $\mathbf{M}$ via $\mathbf{M}_{i,j}$, the j-th column via $\mathbf{M}_j$ and the columns of the matrix indexed by a vector $\mathbf{v}$ via $\mathbf{M}_{\mathbf{v}}$. Substituting $\boldsymbol{\nu}$ into (15) the scaling matrix can be written as

$$\begin{aligned} \mathbf{S}_{i,j} &= \mathbf{K}_{\nu_{i,1},\nu_{i,1}} - \mathbf{K}_{\nu_{i,1},\nu_{j,2}} \\ &\quad -\mathbf{K}_{\nu_{i,2},\nu_{j,1}} + \mathbf{K}_{\nu_{i,2},\nu_{j,2}} \end{aligned} \tag{19}$$

and similarly substituting $\boldsymbol{\nu}$ into (17) and (18) the projected kernel matrix can be expressed as

$$\hat{\mathbf{K}} = \mathbf{K} - (\mathbf{K}_{\boldsymbol{\nu}_1} - \mathbf{K}_{\boldsymbol{\nu}_2}) \mathbf{S}^+ (\mathbf{K}_{\boldsymbol{\nu}_1} - \mathbf{K}_{\boldsymbol{\nu}_2})^T \tag{20}$$

where $\mathbf{K}_{\boldsymbol{\nu}_t} = \left[\mathbf{K}_{\nu_{1,t}} \dots \mathbf{K}_{\nu_{m,t}}\right], t \in \{1, 2\}$.

In our experiments the bandwidth parameter for each point is selected as the k-th smallest distance from the point to all the data points on the feature space, where k is selected as a fraction of the number of total points n. The constrained mean shift algorithm is given in Figure 2.

6. Experiments

We conduct experiments on three datasets. The first set of experiments are performed on challenging synthetic data

Input: Kernel matrix $\mathbf{K}$, Constraint index matrix $\boldsymbol{\nu}$, Bandwidth selection parameter k

- Compute scaling matrix $\mathbf{S}$ via (19), projected kernel matrix $\hat{\mathbf{K}}$ via (20)
- Compute bandwidths h_i as the k-th smallest distance from the point using $\hat{\mathbf{K}}$ and (11), and $d_{\hat{\phi}} = \text{rank}(\hat{\mathbf{K}})$
- Repeat for all data points $i = 1...n$
  - Let $\boldsymbol{\alpha}_{\mathbf{y}_i} = \mathbf{e}_i$
  - Repeat until convergence

$$\bar{\boldsymbol{\alpha}}_i = \frac{\sum_{j=1}^{n} \frac{\mathbf{e}_j}{h_j^{d_{\hat{\phi}}+2}} g\left(\frac{\boldsymbol{\alpha}_{\mathbf{y}_i}^T \hat{\mathbf{K}} \boldsymbol{\alpha}_{\mathbf{y}_i} + \mathbf{e}_j^T \hat{\mathbf{K}} \mathbf{e}_j - 2\boldsymbol{\alpha}_{\mathbf{y}_i}^T \hat{\mathbf{K}} \mathbf{e}_j}{h_i^2}\right)}{\sum_{j=1}^{n} \frac{1}{h_j^{d_{\hat{\phi}}+2}} g\left(\frac{\boldsymbol{\alpha}_{\mathbf{y}_i}^T \hat{\mathbf{K}} \boldsymbol{\alpha}_{\mathbf{y}_i} + \mathbf{e}_j^T \hat{\mathbf{K}} \mathbf{e}_j - 2\boldsymbol{\alpha}_{\mathbf{y}_i}^T \hat{\mathbf{K}} \mathbf{e}_j}{h_j^2}\right)}$$

- Group points $\bar{\boldsymbol{\alpha}}_{\mathbf{y}_i}$ and $\bar{\boldsymbol{\alpha}}_{\mathbf{y}_j}$, $i, j = 1 \ldots n$ satisfying $\bar{\boldsymbol{\alpha}}_i^T \hat{\mathbf{K}} \bar{\boldsymbol{\alpha}}_i + \bar{\boldsymbol{\alpha}}_j^T \hat{\mathbf{K}} \bar{\boldsymbol{\alpha}}_j - 2\bar{\boldsymbol{\alpha}}_i^T \hat{\mathbf{K}} \bar{\boldsymbol{\alpha}}_j = 0$

Figure 2. Constrained Mean Shift Algorithm.

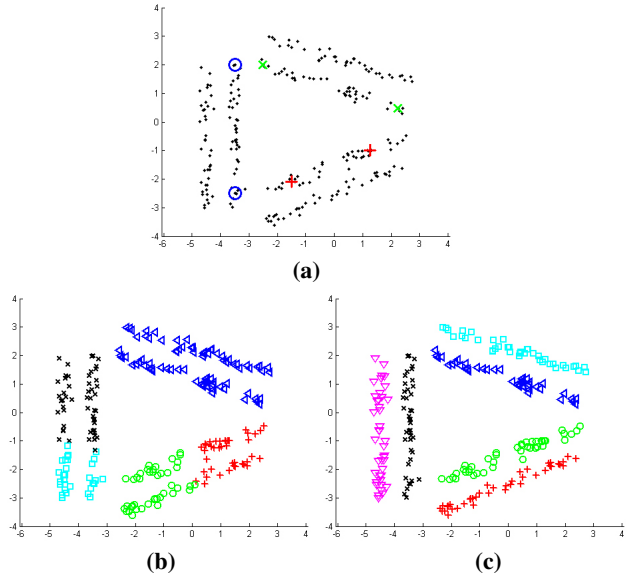

Figure 3. Clustering linear structure. **(a)** Original data. Three constraints are identified with the marked points. **(b)** Mean shift. **(c)** Constrained mean shift.

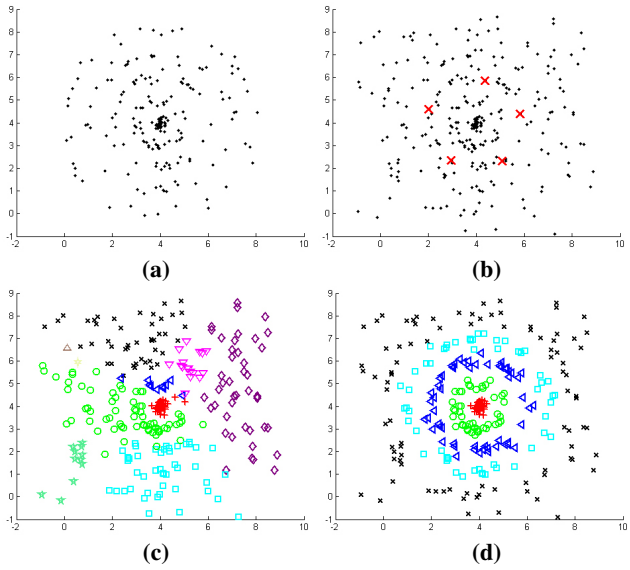

Figure 4. Clustering circular structure. **(a)** Original data. **(b)** Data with outliers. The marked points are identified to be clustered together. **(c)** Mean shift. **(d)** Constrained mean shift.

where the implicit structure of the data are enforced with only a few constraints. In the second experiment, we perform clustering of faces across severe lighting conditions on CMU PIE database [20]. In the third experiment, we perform clustering of object categories on Caltech-4 dataset.

Notice that, our aim in here is not to produce the most accurate clustering given the datasets, therefore we do not focus on selecting the best feature set or modeling the invariances within class. In fact this, would not be a fair comparison since our method has the obvious advantage of being weakly supervised, and generally after introducing a few constraints clustering is almost perfect. We present the accuracy of the approach with respect to the base case which is the kernel mean shift clustering without constraints, and illustrate the effectiveness of the supervision process. We note that the kernel mean shift is also implemented using variable bandwidth and reduces to Euclidean mean shift when a first degree polynomial kernel is used.

In the first two experimental setup we utilize Gaussian kernel $K_N(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}}$ with $\sigma = 5$ for the synthetic experiments and $\sigma = 1000$ for the PIE dataset which are approximately equal to the mean of the pairwise distances between all the points. Note that in general the results are invariant under large perturbations of σ, where larger σ results in smoother cluster boundaries. In the third experiment, we utilize χ^2 kernel, $K_{\chi^2}(\mathbf{x}, \mathbf{x}') = 2\sum_{i=1}^{d} \frac{x(i)x'(i)}{x(i)+x'(i)}$, which is commonly used for histogram based representations.

6.1. Synthetic Experiments

In the first synthetic experiment, we generated 240 data points originating from six different lines in three orientations. Data is corrupted with normally distributed noise with standard deviation 0.1 (Figure 3a). In Figure 3b, we present the result of the mean shift algorithm. Since it has no additional information regarding to the implicit structure of the data points the mean shift algorithm finds five clusters by grouping closer points in the original space. We introduce three constraints shown with the marked points on Figure 3a, where each pair is originated from a line with a different orientation. In Figure 3c, we present the result of constrained mean shift algorithm where the implicit structure of the data is captured with only three constraints, and

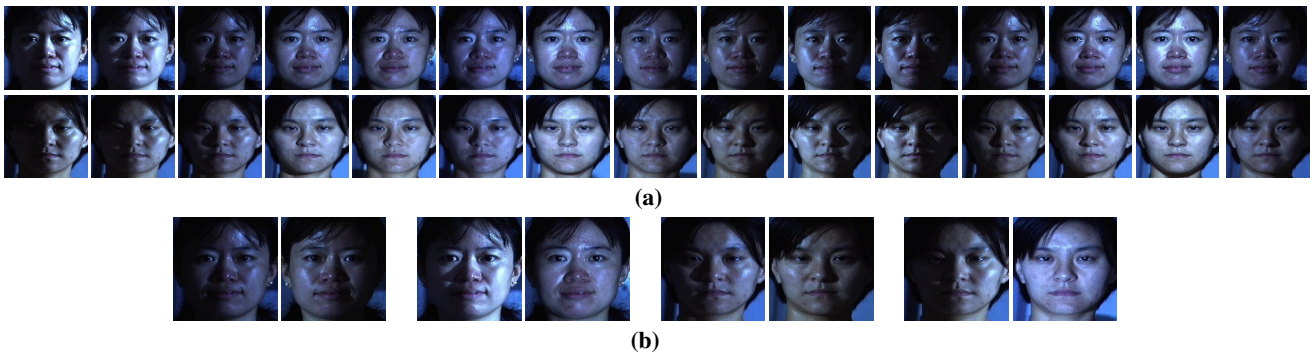

Figure 5. PIE Dataset. **(a)** Fifteen illumination conditions are shown for two subjects. **(b)** Two constraints are specified per subject via pairwise similarities. The nearby samples are constrained to be clustered together.

all the six clusters are perfectly identified.

In the second synthetic experiment, we generated 200 data points originating from five concenteric circles. Data is corrupted with normally distributed noise with standard deviation 0.1 (Figure 4a), and we add 80 outlier points (Figure 4b). This is a very challenging example and it is even difficult to see the circle structure. The mean shift algorithm finds nine clusters by grouping closer points in the original space (Figure 4c). We introduce four constraints enforcing the marked points on Figure 4b to be clustered together. Even though the constraints identify one of the circles, the constrained mean shift algorithm recovers all the five clusters perfectly (Figure 4d). Note that, since we use variable bandwidth, the outliers converge to the nearest mode in the feature space. In both cases, the bandwidth selection parameter is selected as $k = 20$.

6.2. Clustering Faces Across Illumination

The CMU PIE database contains faces of 53 subjects which are imaged under 21 different illumination conditions. We conduct our experiments on a subset of 21 subjects which are selected at random, hence our experimental setup contains 441 images. We coarsely aligned the images with respect to eye and mouth locations and resized them to be 128×128. In Figure 5a, we show 15 illumination conditions for two subjects. Due to significant illumination variation, interclass variability is very large and some of the samples from different subjects appear to be much closer to each other than within classes.

We converted the images to gray scale and normalized between zero and one. Each image is then represented with a 16384-dimensional vector where the columns of the image are concatenated. We note that, better representations can be used by modeling the imaging process or using illumination invariant features, however our motivation in here is to demonstrate the capability of the presented approach for learning the inherent structure of the data.

In this experiment, we consider the scenario where some of the data points can be labeled apriori, and we would like to utilize the prior information for the new observations. To do so, we fix four illumination conditions which are moderately different from each other, and produce two similarity constraints per class. Therefore, 42 constraints are specified for 441 images which is approximately equal to labeling $1/10$ of the dataset. In fact the supervision is weaker in the sense that we do not specify number of subjects or the labels for the constraints. In Figure 5b, four constraints specified for two subjects are shown.

In Figure 6a, we present the pairwise distance matrix (PDM) which is computed on the feature space via (11) using the kernel matrix $\mathbf{K}$, where darker colors indicate smaller distances. The matrix is organized such that the first 21 columns and rows correspond to the first subject and the structure is repeated over the matrix. Ideally this matrix should be block diagonal with 21×21 blocks. Instead, we see that the smaller distances are repeated across the rows and columns which indicate different subjects under same illumination are more closer on the feature space. We perform mean shift clustering using the kernel matrix $\mathbf{K}$ (without nulls pace projection). In Figure 6b, we plot the pairwise distances using the converged modes. The original mean shift clustering finds five modes corresponding to partly illumination conditions partly subject labels and fails to recover the true clusters.

In Figure 6c, we present the PDM after enforcing the pairwise constraints (using the kernel matrix $\hat{\mathbf{K}}$). On this matrix the block diagonal structure is easily observed. In Figure 6d, we plot the pairwise distances using the converged modes recovered through the constrained mean shift algorithm. The algorithm detects all the 21 subjects and clustering accuracy is 100%.

6.3. Clustering Visual Objects

In the third experiment, we perform clustering of visual objects using bag of features representation. We sampled a subset of 400 images from Caltech-4 dataset which includes four object categories: airplanes, cars, faces and motorcycles (Figure 7). From each image we extracted multi scale

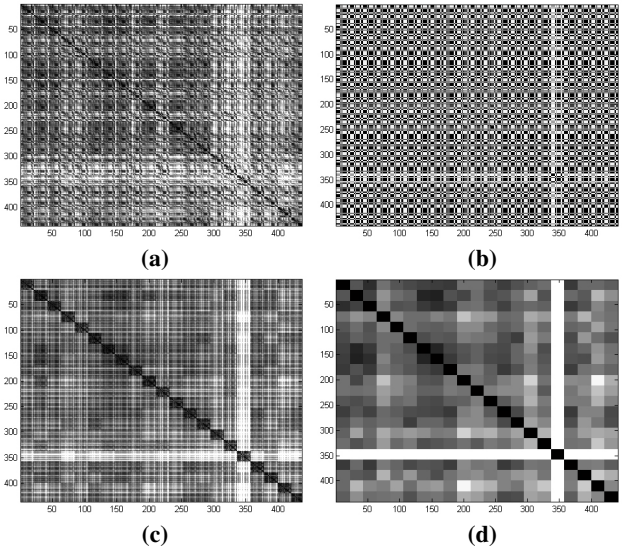

Figure 6. Clustering on PIE dataset. **(a)** Pairwise distance matrix (PDM) on the feature space. The rows and columns are ordered by the class. **(b)** PDM using modes recovered by mean shift. **(c)** PDM after null space projection. **(d)** PDM using modes recovered by constrained mean shift. The accuracy of constrained mean shift is 100% whereas the original mean shift fails to recover the true clusters.

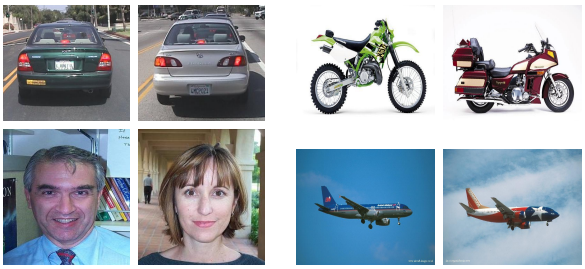

Figure 7. Samples from four categories from Caltech-4 dataset.

DoG interest points and computed SIFT descriptors [19]. There were variable number of feature points per image between 500 and 3500. We generated a vocabulary of 500 visual words by clustering the descriptors via k-means clustering. Each image is then represented with a $d = 500$ bin bag of feature histogram.

In this experiment, we consider the scenario where an expert can guide the clustering process by selecting a few similar examples to be clustered together. Since the selection process is usually at random, we simulate the procedure by enforcing pairwise constraints randomly selected within classes.

In Figure 8a, we present the pairwise distance matrix computed on the feature space induced by the χ^2 kernel. Although the block diagonal structure is more visible with respect to the PIE dataset, there are still confusion within the first and the last clusters which are airplanes and motorcycles classes. We performed the mean shift clustering using the χ^2 kernel and the pairwise distances at the converged modes are shown in Figure 8b. Although the algorithm finds four dominant modes corresponding to four categories, some of the samples from airplanes class and half of the motorcycles class are incorrectly identified as cars (third cluster). The overall clustering accuracy is 74.25% which is computed as the percentage of correct labels after identifying the unique maximal cardinality cluster within each class.

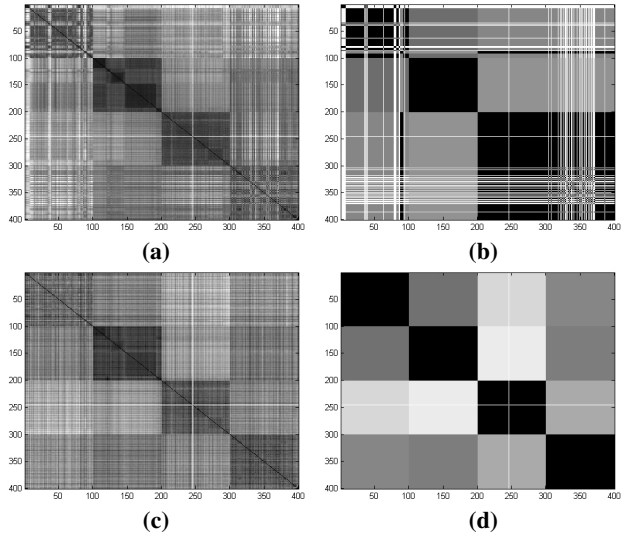

Figure 8. Clustering on Caltech-4 dataset. **(a)** PDM on the feature space. **(b)** PDM using modes recovered through mean shift. Several examples within airplanes and motorcycles class are misclustered into the cars class. **(c)** PDM after enforcing 10 constraints per class. **(d)** PDM using modes recovered through constrained mean shift. Only a single example among 400 is misclustered.

In Figure 8c, we present a typical example of PDM after enforcing 10 pairwise constraints per class which are selected at random. In this matrix, the confusion across the classes are much less and the constrained mean shift algorithm can accurately cluster the objects (Figure 8c). Only a single example among 400 is misclustered.

In Figure 9, we present the clustering error (1-accuracy) with respect to the number of constraints per class. The curve is generated as the average result over 20 runs where at each run a different random set of constraints are selected. We observe that, by introducing only a few constraints the clustering accuracy is significantly improved, and it is over 99% for more than seven constraints per class.

6.4. Complexity

In general, kernel mean shift algorithm performs $20-30$ iterations per data point. Assuming this is a constant factor the computational complexity of the kernel mean shift algorithm is $\mathcal{O}(n^3)$ where n is the number of points. To compute

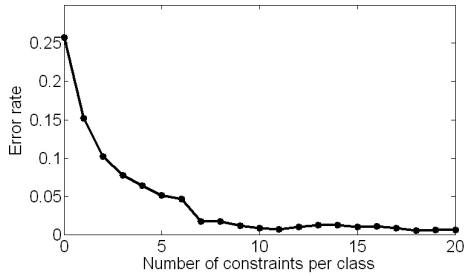

Figure 9. Clustering performance with respect to number of constraints.

the projected kernel matrix for constrained mean shift, additional $\mathcal{O}(m^3 + m^2 n)$ operations are needed where m is the number of constraints. This term is negligible since usually $m << n$. Approximately the algorithm takes 10 seconds for clustering 400 points with a Matlab code.

The bottleneck of the algorithm is the memory requirement since $\mathcal{O}(n^2)$ memory is required for storing kernel matrices. An approximate solution to this problem is using low rank decomposition of kernel matrix using an incremental SVD technique such as [3].

7. Conclusion

We presented a novel constrained mean shift clustering method that can incorporate multiple pairwise must-link priors. Experiments conducted on challenging synthetic and real data show that using only a minimal number of constraints can competently enforce the desired structure on the clustering and drastically improve the performance. The presented approach also extends to inner product spaces thus, it is applicable to a wide range of problems. As a future work, we are going to extend the priors to include the cannot-link pairs and soft similarities.

References

[1] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *Proc. 10th Intl. Conf. on Knowledge Discovery and Data mining*, 2004.

[2] O. Boykov, Y. Veksler and R. Zabih. Fast approximate energy minimisation via graph cuts. In *IEEE Trans. Pat. Anal. Mach. Intell.*, 2001.

[3] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *Proc. European Conf. on Comp. Vis.,* Copehagen, Denmark, pages 707–720, 2002.

[4] M. A. Carreira-Perpinan. Gaussian mean-shift is an EM algorithm. *IEEE Trans. Pat. Anal. Mach. Intell.*, 29(5):767–776, 2007.

[5] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[6] H. Chen and P. Meer. Robust fusion of uncertain information. *IEEE Trans. Sys., Man, Cyber.-Part B*, 35:578–586, 2005.

[7] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pat. Anal. Mach. Intell.*, 17:790–799, 1995.

[8] R. Collins. Mean shift blob tracking through scale space. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Recog.,* Madison, WI, volume 2, pages 234–240, 2003.

[9] D. Comaniciu. Variable bandwidth density-based fusion. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Recog.,* Madison, WI, volume 1, pages 56–66, 2003.

[10] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pat. Anal. Mach. Intell.*, 24:603–619, 2002.

[11] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Recog.,* Hilton Head, SC, volume 1, pages 142–149, 2000.

[12] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[13] M. Fashing and C. Tomasi. Mean shift is a bound optimization. *IEEE Trans. Pat. Anal. Mach. Intell.*, 25:471–474, 2005.

[14] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 1975.

[15] R. Ghani. Combining labeled and unlabeled data for multiclass text categorization. In *Proc. 19th Intl. Conf. on Machine Learning*, 2002.

[16] L. Grady. Random walks for image segmentation. *IEEE Trans. Pat. Anal. Mach. Intell.*, 28(11):1768–1783, 2006.

[17] G. Hager, M. Dewan, and C. Stewart. Multiple kernel tracking with SSD. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Recog.,* Washington, DC, volume 1, pages 790–797, 2004.

[18] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. In *Proc. 22nd Intl. Conf. on Machine Learning*, 2005.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. J. of Comp. Vision*, 60(2):91–110, 2004.

[20] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Trans. Pat. Anal. Mach. Intell.*, 25(12):1615–1618, 2003.

[21] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3D motion estimation via mode finding on Lie groups. In *Proc. 10th Intl. Conf. on Comp. Vis.,* Beijing, China, volume 1, pages 18–25, 2005.

[22] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *Proc. European Conf. on Comp. Vis.,* Marsille, France, pages 705–718, 2008.

[23] K. Wafstaff and S. Rogers. Constrained k-means clustering with background knowledge. In *Proc. 18th Intl. Conf. on Machine Learning*, 2001.

[24] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. In *Proc. European Conf. on Comp. Vis.,* Prague, Czech Republic, volume 2, pages 238–249, 2004.

[25] B. Yan and C. Domeniconi. An adaptive kernel method for semi-supervised clustering. In *Proc. European Conf. on Comp. Vis.,* Graz, Austria, 2006.