Multi-View Anomaly Detection: Neighborhood in Locality Matters[*]

Xiang-Rong Sheng, De-Chuan Zhan, Su Lu, Yuan Jiang
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{shengxr, zhandc, lus, jiangy}@lamda.nju.edu.cn

Abstract

Identifying anomalies in multi-view data is a difficult task due to the complicated data characteristics of anomalies. Specifically, there are two types of anomalies in multi-view data–anomalies that have inconsistent features across multiple views and anomalies that are consistently anomalous in each view. Existing multi-view anomaly detection approaches have some issues, e.g., they assume multiple views of a normal instance share consistent and normal clustering structures while anomaly exhibits anomalous clustering characteristics across multiple views. When there are no clusters in data, it is difficult for existing approaches to detect anomalies. Besides, existing approaches construct a profile of normal instances, then identify instances that do not conform to the normal profile as anomalies. The objective is formulated to profile normal instances, but not to estimate the set of normal instances, which results in sub-optimal detectors. In addition, the model trained to profile normal instances uses the entire dataset including anomalies. However, anomalies could undermine the model, i.e., the model is not robust to anomalies. To address these issues, we propose the nearest neighbor-based MUlti-View Anomaly Detection (MUVAD) approach. Specifically, we first propose an anomaly measurement criterion and utilize this criterion to formulate the objective of MUVAD to estimate the set of normal instances explicitly. We further develop two concrete relaxations for implementing the MUVAD as MUVAD-QPR and MUVAD-FSR. Experimental results validate the superiority of the proposed MUVAD approaches.

Introduction

Anomalies are data patterns that possess different data characteristics from normal instances. Anomaly detection aims at identifying anomalies in a given dataset, which is an important task due to the fact that anomalies often provide significant and critical information. For example, in credit card transactions, anomalies could indicate fraudulent credit card usage (Aleskerov, Freisleben, and Rao 1997). In computer-assisted diagnosis, anomaly detection techniques are widely used to detect anomalous images which could signify the presence of a certain disease (Spence, Parra, and Sajda 2001). Anomaly detection also plays a significant role in network systems, where anomalies could stand for malicious attacks (Ding et al. 2012).

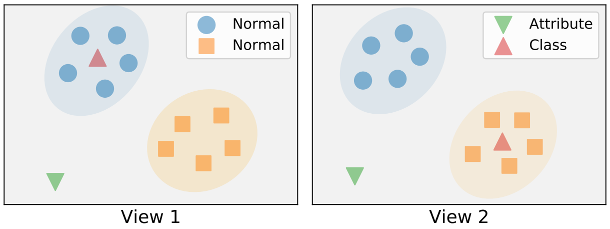


Figure 1: Illustration of normal instances, unanimous anomaly and dissension anomaly. Both views derive from the same original instances. Here, blue circles and orange squares are normal instances with different labels. The red up-triangle is a dissension anomaly and the green down-triangle is an unanimous anomaly. Note that labels of the red up-triangle in two views create dissension and the green down-triangle is consistently distant from normal instances.

Nowadays, data are usually collected from diverse sources and different sources of data exhibit heterogeneous properties. Features from a particular source are regarded as a particular view to describe the object. Multi-view learning approaches can utilize the abundant information of different views and explore the consistency property to get better generalization ability (Blum and Mitchell 1998; Wang and Zhou 2010; Bickel and Scheffer 2004; Jia, Salzmann, and Darrell 2010; Kumar, Rai, and Daumé III 2011; Ye et al. 2015). Although the consistency property provides more information facilitating the discrimination, it also brings more challenges for anomaly detection since anomalies now have more complicated data characteristics. Specifically, anomalies in multi-view data can be classified into two groups: (i) anomaly that has inconsistent features across multiple views and (ii) anomaly that is consistently anomalous in each view. For simplicity, we will refer to view-inconsistent anomaly as **"dissension anomaly"** and consistently anomalous anomaly as **"unanimous anomaly"**. The different characteristics of these two types of anomalies are illustrated in Fig. 1.

[*] This research was supported by the National Key R&D Program of China (2018YFB1004300), the NSFC (61773198, 61632004)

The ability to detect anomalies in multi-view data is a highly desirable feature in many application domains such as micro-expression detection (Yan et al. 2013), purchase behavior analysis (Gao et al. 2011), information disparity management (Duh et al. 2013), malicious insider detection (Liu and Lam 2012), etc. In the task of micro-expression detection, a video clip is given and the goal is to find frames that contain micro-expressions (facial expressions that have short duration and low intensity). Yan et al. (2013) find that unlike conventional facial expressions, micro-expressions usually appear partially (either upper face or lower face). The upper face and lower face are naturally two views to describe the emotion of the subject. Partially appeared micro-expressions are dissension anomalies since the upper face and lower face are inconsistent in expressions and fully appeared micro-expressions are unanimous anomalies since upper face and lower face are consistently anomalous from the neutral facial expressions. Multi-view anomaly detection technique helps the detection of micro-expressions, which benefits various fields such as national security (Ekman 2009).

A straightforward approach for multi-view anomaly detection is to concatenate all views into one single view and adapt the problem to single-view anomaly detection problem. However, this concatenation neglects the abundant and consistent information across multiple views, which results in sub-optimal detectors. Recently, a number of approaches have been proposed for multi-view anomaly detection (Gao et al. 2011; Liu and Lam 2012; Alvarez et al. 2013; Li, Shao, and Fu 2015; Zhao and Fu 2015). Most existing approaches rely on the clustering assumption: multiple views of normal instances share consistent and normal clustering structures while dissension anomalies tend to fall into different clusters and unanimous anomalies consistently deviate from all clusters w.r.t. different views.

We argue that existing multi-view anomaly detection approaches have three main issues: (i) Clustering assumptions on data: when there are no clusters in data, it is difficult for existing approaches to detect anomalies. (ii) Sub-optimal performance due to normal instance profiling: the objective is formulated to profile normal instances, but not formulated to estimate the set of normal instances explicitly. As a consequence, the performance might not be as good as expected. (iii) Lack of robustness: the model is trained with the entire dataset including anomalies. However, anomalies in the dataset could undermine the model, e.g., the centers of clusters might be corrupted by anomalies. Consequently, existing approaches are not robust to anomalies.

To address these issues, we propose the nearest neighbor-based MUlti-View Anomaly Detection (MUVAD) approach that is capable of detecting dissension and unanimous anomalies simultaneously. We outline the major contributions as:

- The MUVAD approach addresses the issue (i) of existing approaches by taking the neighborhood structure of data into account and making no assumption on the clustering structures. Consequently, the MUVAD can handle data that have no clusters.
- We propose a nearest neighbor-based anomaly measurement criterion. In contrast to existing approaches that profile normal instances, we utilize the proposed criterion to formulate the objective which aims at estimating the set of normal instances explicitly. Thus, the MUVAD can solve the issue (ii) of existing approaches.
- We develop two concrete relaxations for implementing the MUVAD as MUVAD-QPR and MUVAD-FSR. By downweighting anomalies, the MUVAD mitigate the effects of anomalies and improve the robustness. Thus, the MUVAD can solve the issue (iii) of existing approaches.
- We apply the MUVAD approach on datasets from different domains. Experimental results demonstrate the superiority of the MUVAD approach.

The rest of paper starts with a detailed description of the methodology of the MUVAD. Then two concrete relaxations are developed for implementing the MUVAD. Next is a review of existing multi-view anomaly detection approaches, followed by empirical evaluation and conclusion.

Methodology

This section first clarifies notations and gives formal definitions of two types of multi-view anomalies and t-nearest $\mathcal{C}$ neighbors. Then a novel anomaly measurement criterion is proposed based on these definitions. The objective is further formulated to estimate the set of normal instances explicitly.

Preliminaries

Assume we are handling a dataset $\mathcal{D}$ of N instances and V views, $\mathcal{D} = \{X_i | i = 1, 2, \ldots, N\}$, where $X_i = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^V)$ is the i-th instance. $\mathbf{x}_i^v$ denotes the v-th view of X_i. Assume $\mathcal{D}$ consists of N_0 normal instances and $N - N_0$ anomalies. The $N - N_0$ anomalies are either dissension or unanimous anomalies. Let $\mathcal{S}$ denotes the set of all normal instances' indexes, i.e., $i \in \mathcal{S}$ represents X_i is a normal instance. Let $\odot$ denotes the element-wise product operator. Let $\text{card}(\mathcal{C})$ denotes the cardinality of set $\mathcal{C}$. $|\boldsymbol{v}|$ represents taking absolute value of each element in vector $\boldsymbol{v}$. $\mathbf{1}$ and $\mathbf{0}$ are column vectors with all of the elements equaling to 1 and 0, respectively. $\text{diag}(A)$ represents taking diagonal elements of matrix A. Let K^v denotes the similarity matrix in the v-th view, where $K_{ij}^v \geq 0$ measures the similarity between $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$. K_i^v represents the i-th row of K^v. To make nearest neighbor selection unique and repeatable, we will add a small random positive value to elements in K_i^v that are equal in value.

We define dissension and unanimous anomaly as:

Definition 1. ***Dissension anomaly*** *is an anomaly that possesses inconsistent characteristics (e.g., different class membership) across different views.*

Definition 2. ***Unanimous anomaly*** *is an anomaly that possesses consistent anomalous characteristics in each view.*

We define t-nearest $\mathcal{C}$ neighbors as:

Definition 3. *Let* $\mathcal{C} = \{c_1, \ldots, c_k, \ldots, c_{N_0}\}$ *denotes an index set of size* N_0*, where* $1 \leq c_k \leq N$ *and* $c_k \in \mathbb{Z}$*. Let*

*$\mathcal{D}^{\mathcal{C}} = \{X_k | k \in \mathcal{C}\}$ denotes the dataset that consists of instances with indexes in $\mathcal{C}$. For any instance X_i in $\mathcal{D}$, its **t-Nearest $\mathcal{C}$ Neighbors** in the v-th view is defined as the set of t instances in $\mathcal{D}^{\mathcal{C}}$ (not including i if $i \in \mathcal{C}$) that are most similar to $\mathbf{x}_i^v$ in the v-th view.*

Let $\mathcal{N}_{\mathcal{C}}^t(\mathbf{x}_i^v)$ denotes the t-nearest $\mathcal{C}$ neighbors of $\mathbf{x}_i$ in the v-th view. Note that when $\mathcal{C} = \mathcal{S}$, then all instances' t-nearest $\mathcal{C}$ neighbors are normal instances. We will refer to $\mathcal{N}_{\mathcal{S}}^t(\mathbf{x}_i^v)$ as ***t*-Nearest Normal Neighbors** of $\mathbf{x}_i^v$ in the v-th view. Take Fig. 1 as an example: for the up-triangle dissension anomaly, its 5-nearest normal neighbors in the first view are the five circle normal instances and 5-nearest normal neighbors in the second view are the five square normal instances.

To simplify notations and explanations, we will start with the case where $\mathcal{D}$ has two views and will further extend to a more general case where $\mathcal{D}$ has multiple views.

Anomaly Measurement Criterion

Since multi-view data provide abundant and consistent information, it's natural to assume multiple views of a normal instance have similar neighborhood structures, i.e., for a normal instance, its t-nearest normal neighbors in one view will also be similar to it in the other view. Thus, we propose a nearest neighbor-based anomaly measurement criterion as:

$$s^t(X_i) = \sum_{j \in \mathcal{N}_{\mathcal{S}}^t(\mathbf{x}_i^1)} K_{ij}^2 + \sum_{j \in \mathcal{N}_{\mathcal{S}}^t(\mathbf{x}_i^2)} K_{ij}^1.$$

This criterion helps identify dissension and unanimous anomalies simultaneously. The interpretations are given as follows:

- For a normal instance X_i, since it is consistent across multiple view, its t-nearest normal neighbors in one view should be similar to it in the other view, which would give rise to a large value of $s^t(X_i)$.
- For a dissension anomaly X_i, since it is inconsistent across multiple views, its t-nearest normal neighbors in one view could be dissimilar to it in the other view, leading to a small value of $s^t(X_i)$.
- For an unanimous anomaly X_i, since it is consistently anomalous in each view, it is dissimilar to normal instances in each view. Thus, its t-nearest normal neighbors in one view are also dissimilar to it in the other view, resulting in a small value of $s^t(X_i)$ as well.

Thus, both dissension and unanimous anomalies could result in small anomaly measurement values. We will refer to this criterion as ***s*-Score** in the following. Note that although simple and effective, s-score can not be used directly in practice since the computation of s-score uses the set of normal instances, which is unknown beforehand.

The Formulation

Given an index set $\mathcal{C}$ of size N_0, we define ***u*-Score** for X_i w.r.t. $\mathcal{C}$ as:

$$u_{\mathcal{C}}^t(X_i) = \sum_{j \in \mathcal{N}_{\mathcal{C}}^t(\mathbf{x}_i^1)} K_{ij}^2 + \sum_{j \in \mathcal{N}_{\mathcal{C}}^t(\mathbf{x}_i^2)} K_{ij}^1.$$

Note that when $\mathcal{C} = \mathcal{S}$, $u_{\mathcal{C}}^t(X_i) = s^t(X_i)$. We estimate the set of normal instances and $\boldsymbol{s}$-scores by formulating the objective as:

$$\begin{aligned} \max_{\mathcal{C}} \quad & \sum_{i \in \mathcal{C}} u_{\mathcal{C}}^t(X_i) \\ \text{s.t.} \quad & \text{card}(\mathcal{C}) = N_0,\ \mathcal{C} = \{y | y \in \mathbb{Z}, 1 \leq y \leq N\}. \end{aligned} \tag{1}$$

The optimal solution to Eq. 1 is the estimated index set of normal instances. Eq. 1 means that we should find a set $\mathcal{C}$ that the sum of $\boldsymbol{u}$-scores of $\mathcal{C}$ is largest, where the $\boldsymbol{u}$-scores are calculated using t-nearest $\mathcal{C}$ neighbors. This is reasonable since the $\boldsymbol{u}$-score of an anomaly in $\mathcal{C}$ would be smaller than the $\boldsymbol{s}$-score of a normal instance that are not in $\mathcal{C}$. Besides, anomalies in $\mathcal{C}$ that appear as other normal instance's t-nearest $\mathcal{C}$ neighbors would make the $\boldsymbol{u}$-scores of these normal instances be smaller than their $\boldsymbol{s}$-scores.

In order to simplify the optimization problem in Eq. 1, we introduce auxiliary variables $\boldsymbol{O}$ and $W = \{W^1, W^2\}$, where $\boldsymbol{O} \in \{0,1\}^N$ and $W^1, W^2 \in \{0,1\}^{N \times N}$. $\boldsymbol{O}$ has N_0 elements that equal to 1. Note that given $\boldsymbol{O}$, there exists a corresponding index set $\mathcal{C}_{\boldsymbol{O}}$ of size N_0, where $O_i = 1$ and $O_i = 0$ represents that i is in $\mathcal{C}_{\boldsymbol{O}}$ and not in $\mathcal{C}_{\boldsymbol{O}}$, respectively. Each row of W^1, W^2 has t elements that equal to 1 and the diagonal elements of W^1, W^2 equal to 0. Eq. 1 can be equivalently rewritten as:

$$\begin{aligned} \max_{\boldsymbol{O}, W} \quad & \ell(\boldsymbol{O}, W) = \sum_{i,j}^{N} O_i O_j W_{ij}^1 K_{ij}^2 + \sum_{i,j}^{N} O_i O_j W_{ij}^2 K_{ij}^1 \\ \text{s.t.} \quad & W^v \in \{0,1\}^{N \times N},\ W^v \mathbf{1} = t\mathbf{1},\ \text{diag}(W^v) = \mathbf{0}, \\ & \min_{j \in \{k : W_{ik}^v = 1\}} O_j K_{ij}^v > \max_{j \in \{k : W_{ik}^v = 0 \wedge k \neq i\}} O_j K_{ij}^v, \\ & \boldsymbol{O} \in \{0,1\}^N,\ \boldsymbol{O}^\top \mathbf{1} = N_0, \\ & \forall v \in \{1,2\},\ i \in \{1, \ldots, N\}. \end{aligned} \tag{2}$$

Proposition 1. *Let $\boldsymbol{O}^\star, W^\star$ denote the optimal solution to Eq. 2 and $\mathcal{C}_{\boldsymbol{O}^\star}$ denotes the corresponding index set, respectively. Then $\mathcal{C}_{\boldsymbol{O}^\star}$ is the optimal solution to Eq. 1.*

Proof. Note that

$$\max_{\boldsymbol{O}, W} \ell(\boldsymbol{O}, W) = \max_{\mathcal{C}_{\boldsymbol{O}}} \max_{W} \ell(\mathcal{C}_{\boldsymbol{O}}, W).$$

Given $\mathcal{C}_{\boldsymbol{O}}$, the optimization problem over W can be solved by optimizing $W_1^1, \ldots, W_N^1, W_1^2, \ldots, W_N^2$ separately. Due to the constraints w.r.t. W_i^v, the solution $W_i^{v\star}$ is the indicator vector of $\mathcal{N}_{\mathcal{C}_{\boldsymbol{O}}}^t(\mathbf{x}_i^v)$, i.e., $W_{ij}^v = 0$ for $j \notin \mathcal{N}_{\mathcal{C}_{\boldsymbol{O}}}^t(\mathbf{x}_i^v)$ and $W_{ij}^v = 1$ for $j \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{O}}}^t(\mathbf{x}_i^v)$. Thus, we can get:

$$\begin{aligned} & \max_{\mathcal{C}_{\boldsymbol{O}}} \max_{W} \ell(\mathcal{C}_{\boldsymbol{O}}, W) \\ & = \max_{\mathcal{C}_{\boldsymbol{O}}} \sum_{i \in \mathcal{C}_{\boldsymbol{O}}} \Big(\sum_{j \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{O}}}^t(\mathbf{x}_i^1)} K_{ij}^2 + \sum_{j \in \mathcal{N}_{\mathcal{C}_{\boldsymbol{O}}}^t(\mathbf{x}_i^2)} K_{ij}^1 \Big) \\ & = \max_{\mathcal{C}_{\boldsymbol{O}}} \sum_{i \in \mathcal{C}_{\boldsymbol{O}}} u_{\mathcal{C}_{\boldsymbol{O}}}^t(X_i). \end{aligned}$$

Thus, $\mathcal{C}_{\boldsymbol{O}^\star}$ is the optimal solution to Eq. 1. □

The objective $\ell(\boldsymbol{O}, W)$ in Eq. 2 can be rewritten in matrix form as:

$$\ell(\boldsymbol{O}, W) = \boldsymbol{O}^\top (K^2 \otimes W^1 + K^1 \otimes W^2)\boldsymbol{O}. \tag{3}$$

The discussion for two-view data can be naturally adapted to V-view data ($V \geq 2$) by considering the combination of every two views. The formulation for V-view data is:

$$\begin{aligned} \max_{W,\boldsymbol{O}} \quad & \boldsymbol{O}^\top (\sum_{i \neq j}^{V} K^i \otimes W^j)\boldsymbol{O} \\ \text{s.t.} \quad & W^v \in \{0,1\}^{N \times N},\ W^v \mathbf{1} = t\mathbf{1},\ \mathrm{diag}(W^v) = \mathbf{0}, \\ & \min_{j \in \{k: W^v_{ik}=1\}} O_j K^v_{ij} > \max_{j \in \{k: W^v_{ik}=0 \wedge k \neq i\}} O_j K^v_{ij}, \\ & \boldsymbol{O} \in \{0,1\}^N,\ \boldsymbol{O}^\top \mathbf{1} = N_0, \\ & \forall v \in \{1, \dots, V\},\ i \in \{1, \dots, N\}. \end{aligned} \tag{4}$$

As a matter of fact, $\boldsymbol{O}$ can be regarded as a weights vector, where each element O_i corresponds to a nonnegative weight for X_i. The idea is then to preserve normal instance by assigning its weight to 1 and downweight anomaly by assigning its weight to 0. N_0 can be treated as a hyperparameter. Note that now we can only select a fix number of anomalies each time and cannot sort all instances according to their intensity of anomaly. In the next section, we will develop two relaxation approaches which provide a way to choose the number of anomalies adaptively based on the value of $\boldsymbol{O}$.

Optimization

Since we only care about the relative magnitude of $\boldsymbol{O}$, it is more natural and flexible to use softer weight rather than hard weight $\{0,1\}$, which also provides a way to choose the number of anomalies adaptively. Thus, we relax the integer constraint of $\boldsymbol{O}$ in Eq. 4 in two different ways and develop corresponding approaches for optimization. The basic idea behind these two approaches is to iteratively optimize $\boldsymbol{O}$ and then update W to satisfy the constraints. When the variation of objective value is smaller than the predefined threshold, the algorithms terminate with the current solution. The two proposed relaxation approaches are quadratic programming relaxation (MUVAD-QPR) approach and fast spectral relaxation (MUVAD-FSR) approach.

Quadratic Programming Relaxation

To simplify notations, let $A^W = \sum_{i \neq j}^{V} K^i \otimes W^j$. Since $\mathrm{diag}(W^v) = \mathbf{0}$ for all $v \in \{1, \dots, V\}$, A^W is an indefinite matrix with diagonal elements that equal to 0. Since $O_i \in \{0,1\}$ and $\boldsymbol{O}^\top \mathbf{1} = N_0$, the objective $\boldsymbol{O}^T A^W \boldsymbol{O}$ can be replaced with:

$$\boldsymbol{O}^\top B^W \boldsymbol{O}, \tag{5}$$

where $B^W = A^W + \lambda I$. I is a $N \times N$ identity matrix and $\lambda < 0$. So far, the optimal solution remains the same as in Eq. 4. We relax the integer constraint of O_i to $O_i \in [0,1]$ for all $i \in \{1, \dots, N\}$ and solve the problem in an alternative manner:

Optimizing $\boldsymbol{O}$ when W is fixed: When W is fixed, the sub-problem becomes:

$$\max_{\boldsymbol{O}} \boldsymbol{O}^\top C^W \boldsymbol{O},\ \text{s.t.}\ \mathbf{0} \leq \boldsymbol{O} \leq \mathbf{1},\ \boldsymbol{O}^\top \mathbf{1} = N_0, \tag{6}$$

where $C^W = (B^W + (B^W)^\top)/2$. Note that when λ is taken a small value, C^W is guaranteed to be negative definite. The sub-problem can be converted to a convex quadratic program by minimizing the negative of the objective, which can be efficiently solved by off-the-shelf solvers, e.g., quadprog in MATLAB.

Updating W when $\boldsymbol{O}$ is fixed: When $\boldsymbol{O}$ is fixed, $W^1_1, \dots, W^1_N, \dots, W^v_1, \dots, W^v_N, \dots, W^V_1, \dots, W^V_N$ can be updated separately. Since W^v_i satisfies $W^v_i \in \{0,1\}^N$, $W^v_i \mathbf{1} = t$, $W^v_{ii} = 0$ and

$$\min_{j \in \{k: W^v_{ik}=1\}} O_j K^v_{ij} \geq \max_{j \in \{k: W^v_{ik}=0 \wedge k \neq i\}} O_j K^v_{ij},$$

the solution $W^{v\star}_i$ can be obtained by t-nearest neighbor search of X_i based on the weighted similarity. The weighted similarity between X_i and X_j is $O_j K^v_{ij}$. $W^{v\star}_{ij} = 1$ if and only if j is one of the t-nearest neighbors of X_i.

Fast Spectral Relaxation

Since $O_i \in \{0,1\}$ and $\boldsymbol{O}^\top \mathbf{1} = N_0$, the objective $\boldsymbol{O}^T A^W \boldsymbol{O}$ can be replaced with:

$$\boldsymbol{O}^\top (A^W + \gamma \mathbf{1}\mathbf{1}^T)\boldsymbol{O}, \tag{7}$$

where $\gamma > 0$. So far, the optimal solution remains the same as in Eq. 4. The integer constraint $O_i \in \{0,1\}$ can be rewritten as $O_i^2 - O_i = 0$. By summarizing all $N+1$ constraints w.r.t. $\boldsymbol{O}$:

$$O_i^2 - O_i = 0,\ \sum_{i=1}^{N} O_i = N_0,$$

we can get a spectral relaxation problem:

$$\begin{aligned} \max_{W,\boldsymbol{O}} \quad & \boldsymbol{O}^\top (A^W + \gamma \mathbf{1}\mathbf{1}^T)\boldsymbol{O} \\ \text{s.t.} \quad & W^v \in \{0,1\}^{N \times N},\ W^v \mathbf{1} = t\mathbf{1},\ \mathrm{diag}(W^v) = \mathbf{0}, \\ & \min_{j \in \{k: W^v_{ik}=1\}} O_j K^v_{ij} > \max_{j \in \{k: W^v_{ik}=0 \wedge k \neq i\}} O_j K^v_{ij}, \\ & \|\boldsymbol{O}\|_2^2 = N_0, \forall v \in \{1, \dots, V\},\ i \in \{1, \dots, N\}. \end{aligned} \tag{8}$$

Similarly, we solve the optimization problem in Eq. 8 in an alternative manner:

Optimizing $\boldsymbol{O}$ when W is fixed: When W is fixed, $A^W + \gamma \mathbf{1}\mathbf{1}^T$ can be replaced with a symmetric matrix D^W, where $D^W = (A^W + (A^W)^\top)/2 + \gamma \mathbf{1}\mathbf{1}^T$. Thus, the sub-problem becomes:

$$\max_{\boldsymbol{O}} \boldsymbol{O}^\top D^W \boldsymbol{O},\ \text{s.t.}\ \|\boldsymbol{O}\|_2^2 = N_0. \tag{9}$$

Let $\boldsymbol{v}$ denotes the leading eigenvector of D^W, i.e., the one associated with the largest eigenvalue. It can be proved that $\sqrt{N_0}|\boldsymbol{v}|$ is one of the optimal solutions to Eq. 9.

Proposition 2. *$\sqrt{N_0}|\boldsymbol{v}|$ is one of the optimal solutions to Eq. 9*

Proof. Since $\boldsymbol{v}$ is the first eigenvector of D^W, $\sqrt{N_0}\boldsymbol{v}$ is one of the optimal solution to Eq. 9. Replacing $\boldsymbol{O}$ with $\sqrt{N_0}|\boldsymbol{v}|$ in the objective of Eq. 9, we can get:

$$N_0|\boldsymbol{v}^\top|D^W|\boldsymbol{v}| = N_0|\boldsymbol{v}^\top D^W \boldsymbol{v}| \geq N_0 \boldsymbol{v}^\top D^W \boldsymbol{v}.$$

The first equality comes from the fact that all elements in D^w are nonnegative. Thus, $\sqrt{N_0}|\boldsymbol{v}|$ is one of the optimal solutions to Eq. 9. □

Updating W when $\boldsymbol{O}$ is fixed: When $\boldsymbol{O}$ is fixed, the update of W is as same as in quadratic programming relaxation approach.

The effect of $\gamma\mathbf{1}\mathbf{1}^T$ is to regularize $\boldsymbol{O}$ to be smooth and prevent the weights from concentrating on only a few instances. Note that in fast spectral relaxation approach, the solution is irrelevant to N_0 since we only care about the relative magnitude of $\boldsymbol{O}$. Calculating the leading eigenvector of a matrix can be solved efficiently by existing methods, e.g., power method. Each iteration in power method takes $O(N^2)$, which is more efficient compared with the cubic complexity of convex quadratic programming.

The intuition behind the two relaxation approaches is iteratively reweighting data and refining the estimated t-nearest normal neighbors. Although $\boldsymbol{O}$ can no longer be used to indicate anomalies after the relaxation, we find that the relative magnitude of $\boldsymbol{O}$ are well kept after the relaxation, i.e., a normal instance X_i is more likely to have a larger weight O_i. Since the relaxation approaches downweight anomalies, they are also robust to anomalies. When stop criterion meets, we can sort $\boldsymbol{O}$ in ascending order and use a cut-off threshold to select anomalies.

Related Work

Anomaly detection is an important research topic in pattern recognition and data mining (Chandola, Banerjee, and Kumar 2009; Akoglu, Tong, and Koutra 2015). The importance of anomaly detection is due to the fact that anomalies in data often provide significant and critical information. Various approaches have been proposed for anomaly detection (Breunig et al. 2000; Schölkopf et al. 2001; Bay and Schwabacher 2003; Liu, Ting, and Zhou 2008). However, most existing anomaly detection approaches focus on single-view data.

There has been great interest in multi-view learning (Xu, Tao, and Xu 2013). Anomaly detection for multi-view data is a new research topic. The pioneering work on this topic is horizontal anomaly detection (HOAD) approach (Gao et al. 2011). HOAD first constructs a combined similarity graph of each view. Then the k smallest eigenvectors of the graph Laplacian are calculated as spectral embeddings of the instances. The anomalous score is defined as the cosine distance between spectral embeddings of different views. HOAD can be regarded as performing constrained spectral clustering in each view firstly and then finding instances that belong to different clusters in different views. Liu and Lam (2012) proposed a multi-view anomaly detection approach using consensus clustering (CC). CC also aims at detecting dissension anomalies by exploring the inconsistency of clustering results across multiple views. It is worth noting that both HOAD and CC are only designed to detect dissension anomalies. To detect both dissension anomalies and unanimous anomalies simultaneously, Li, Shao, and Fu (2015) proposed multi-view low-rank analysis (MLRA). MLRA performs cross-view low-rank analysis to reveal the intrinsic structures of data. To detect two types of anomalies simultaneously, they design a criterion to estimate the anomalous scores by analyzing the obtained representation coefficients. Zhao and Fu (2015) proposed Dual-Regularized Multi-View Outlier Detection (DMOD) for detecting two types of anomalies simultaneously. DMOD represents multi-view data with latent coefficients and sample-specific errors and characterize each kind of anomaly explicitly. An anomaly measurement function is designed to detect both dissension and unanimous anomalies jointly.

Experiment

To evaluate the proposed MUVAD approaches, we perform experiments on synthetic datasets that have no clusters, benchmark datasets and real world multi-view anomaly detection task. We compare the proposed MUVAD-QPR and MUVAD-FSR approaches with OCSVM (Schölkopf et al. 2001), HOAD (Gao et al. 2011), CC (Liu and Lam 2012), MLRA (Li, Shao, and Fu 2015), DMOD (Zhao and Fu 2015) and CRMOD (Zhao et al. 2018). CRMOD is the extended version of DMOD. Notably, OCSVM is a representative approach for single-view anomaly detection and we include it to investigate the performance of single-view approach on multi-view data. As for OCSVM, multiple views are first concatenated into one single view and then used as input. As for MUVAD-QPR, we use $t = 7, N_0 = 0.9N, \lambda = -2000$ as default parameters. As for MUVAD-FSR, we use $t = 7, \gamma = 2000$ as default parameters. We use the area under the ROC curve (AUC) as the evaluation measure. The higher the AUC is, the better the approach performs.

Synthetic Data

This experiment is performed on synthetic data that have no clusters. We simulate two-view datasets of size $N = 400$ with 398 normal instances, 1 dissension anomaly and 1 unanimous anomaly. Both of the two views X^1, X^2 have 2 features. For normal instances and dissension anomaly, the first view is sampled from the uniform distribution:

$$\mathbf{x} \sim \text{Uniform}(\mathbf{x}|0.9 \leq \|\mathbf{x}\| \leq 1).$$

The first view of unanimous anomaly is sampled from:

$$\mathbf{x} \sim \text{Uniform}(\mathbf{x}|0.4 \leq \|\mathbf{x}\| \leq 0.5).$$

We perform kernel PCA on X^1 with RBF kernel and keep all components. Let Z denotes the projection by kernel PCA. Each instance $\mathbf{x}_i$ corresponds to projection data $\mathbf{z}_i$. For normal instance and unanimous anomaly, we set their second view to the projected data. The unanimous anomaly is consistently anomalous in each view. For dissension anomaly X_i, we generate inconsistent view by setting the second view to $-\mathbf{z}_i$. We repeat the generation procedure for 50 times. Fig. 2 illustrates one of the synthetic datasets. We evaluate all approaches on the synthetic data with default parameters. The results are reported in Tab. 1.

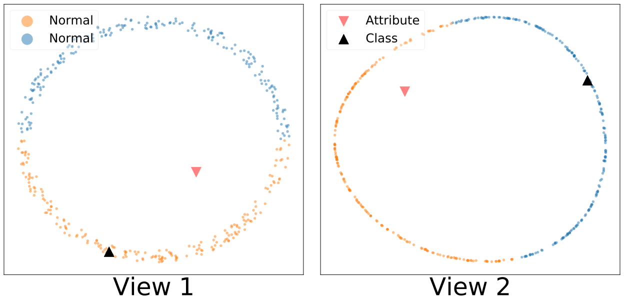


Figure 2: Illustration of a multi-view dataset that has no clusters. Both views derive from the same original instances. Here, blue and orange circles represent normal instances. Red down-triangle represents unanimous anomaly and black up-triangle represents dissension anomaly.

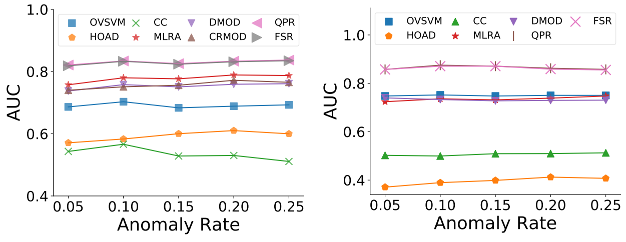
Table 1: Comparison on synthetic data. The mean AUC ± std. are shown in the table.

	AUC (mean ± std)
OCSVM	0.500 ± 0.000
HOAD	0.806 ± 0.045
CC	0.519 ± 0.175
MLRA	0.505 ± 0.193
DMOD	0.184 ± 0.121
CRMOD	0.196 ± 0.121
MUVAD-QPR	**1.000 ± 0.000**
MUVAD-FSR	**1.000 ± 0.000**

From the results, we can see that both MUVAD-QPR and MUVAD-FSR achieve highest performance. The performance of OCSVM is not good and the std. is 0. We observe that OCSVM always puts the unanimous anomaly at the first position and the dissension anomaly at the last position in the ranked list. This is due to the fact that OCSVM is a single-view anomaly detection approach and is difficult to identify dissension anomaly. The AUCs of HOAD, CC, MLRA, DMOD, CRMOD are lower than MUVAD-QPR and MUVAD-FSR, because they rely on the clustering assumption and are difficult to identify anomalies when there are no clusters in data.

Benchmark Dataset

This experiment compares the proposed approaches with others on benchmark datasets from UCI Machine Learning Repository[1] and real world multi-view applications. Following (Gao et al. 2011; Alvarez et al. 2013; Li, Shao, and Fu 2015; Zhao and Fu 2015), we employ three UCI datasets, Ionosphere, Vowel and Zoo for comparison. Two-views are generated by splitting features into two subsets, where each subset corresponds to one view of the data. For each dataset, we also strictly follow (Zhao and Fu 2015) to generate anomalies: for dissension anomaly, we randomly sample two instances X_i, X_j from different classes and swap their first view $\mathbf{x}_i^1, \mathbf{x}_j^1$; for unanimous anomaly, we random sample an instance and then replace its features with random values. To evaluate the performance on dataset with more than two views, we also include two real world multi-view datasets, NewsM and NewsNG, which are extracted from the 20 Newsgroup datasets and have 3 views (Bisson and Grimal 2012). The same anomaly generation procedure is performed on two chosen views. For each dataset, three subsettings are considered: (i) 2% dissension anomalies + 8% unanimous anomalies; (ii) 5% dissension anomalies + 5% unanimous anomalies; (iii) 8% dissension anomalies + 2% unanimous anomalies. For each dataset and each subsetting, we repeat the anomaly generation procedure for 50 times. All anomaly detection approaches are then evaluated on these datasets and the results are reported in Tab. 2.



(a) Ionosphere

(b) Vowel

Figure 3: Changes of AUC as the number of anomalies increases.

From Tab. 2, we can see that the proposed approaches alomost consistently achieve the highest performance. The above observation is within expectation since the objective in Eq. 4 is formulated to estimate the set of normal instances explicitly, while other multi-view anomaly detectors are formulated to profile normal instances and single-view anomaly detection approach can not utilize the full information of multi-view data to improve its performance.

The changes of AUC given different amount of anomalies are also investigated, as shown in Fig. 3. We increase the number of anomalies from 5% to 25%, where the number of dissension and unanimous anomaly are same. Results on two datasets are listed. The performance of MUVAD-QPR and MUVAD-FSR are on par with each other. Besides, MUVAD-QPR and MUVAD-FSR consistently achieve highest performance with all the anomaly ratios.

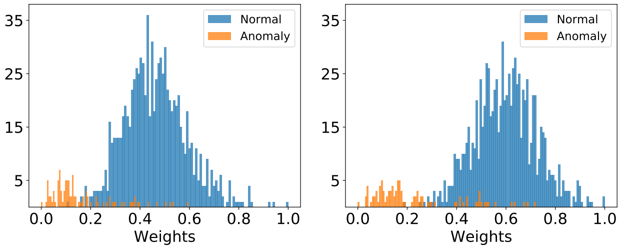
It is beneficial to understand how our reweighting approaches contribute to the anomaly detection. We measure the weights distribution $\boldsymbol{O}$ of instances. The weights are normalized and results on Vowel dataset are listed. As shown in Fig. 4, MUVAD-FSR and MUVAD-QPR assign smaller weights to most anomalies and this makes the proposed approaches detect anomalies reliably.

The efficiency of MUVAD is also investigated. Convergence trends on Ionosphere dataset are illustrated. As shown in Fig. 5, the objectives converge at a small number of iteration, both less than 5 and MUVAD-FSR is less than MUVAD-QPR. Since MUVAD-FSR also has lower computational complexity per iteration, MUVAD-FSR is faster

[1] http://archive.ics.uci.edu/ml

Table 2: Comparison on benchmark datasets. The setting is formatted as "Name-Number of view-Dissension anomaly ratio (%)-Unanimous anomaly ratio (%)". The last two rows list the number of times MuVAD approaches W/T/L (win/tie/loss) when compared with other approaches (pairwise t-tests at 95% significance level). The remaining rows list the mean AUC±std. on the corresponding datasets.

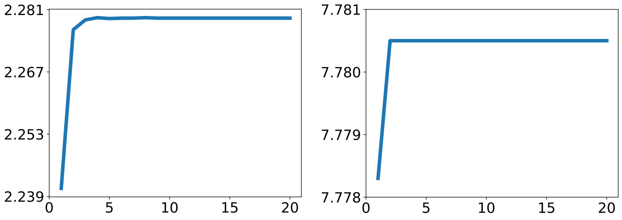
	MuVAD-QPR	MuVAD-FSR	OCSVM	HOAD	CC	MLRA	DMOD	CRMOD
Ionosphere-2-2-8	**0.834±0.013**	0.833±0.014	0.725±0.029	0.655±0.051	0.584±0.080	0.831±0.022	0.741±0.023	0.833±0.033
Ionosphere-2-5-5	**0.834±0.018**	0.833±0.018	0.703±0.045	0.583±0.057	0.566±0.071	0.780±0.022	0.758±0.030	0.743±0.026
Ionosphere-2-8-2	**0.809±0.021**	0.805±0.022	0.642±0.047	0.547±0.064	0.539±0.052	0.714±0.023	0.766±0.033	0.765±0.030
Vowel-2-2-8	**0.886±0.005**	0.879±0.012	0.879±0.015	0.349±0.030	0.499±0.025	0.735±0.028	0.851±0.017	0.833±0.021
Vowel-2-5-5	**0.875±0.012**	0.871±0.009	0.755±0.025	0.389±0.035	0.500±0.033	0.736±0.021	0.733±0.021	0.717±0.027
Vowel-2-8-2	**0.865±0.012**	0.862±0.010	0.626±0.032	0.454±0.029	0.499±0.028	0.739±0.035	0.619±0.022	0.584±0.031
Zoo-2-2-8	**0.866±0.031**	**0.866±0.031**	0.445±0.076	0.491±0.076	0.488±0.066	0.510±0.048	0.823±0.036	0.521±0.071
Zoo-2-5-5	**0.891±0.037**	**0.891±0.037**	0.500±0.109	0.474±0.101	0.487±0.087	0.524±0.088	0.786±0.043	0.496±0.102
Zoo-2-8-2	**0.908±0.030**	**0.908±0.031**	0.488±0.085	0.525±0.085	0.523±0.105	0.532±0.078	0.727±0.055	0.496±0.079
NewsM-3-2-8	0.892±0.016	**0.896±0.019**	0.854±0.015	0.498±0.003	0.474±0.037	0.743±0.034	0.873±0.017	0.883±0.017
NewsM-3-5-5	0.736±0.036	**0.741±0.039**	0.707±0.036	0.548±0.033	0.504±0.042	0.649±0.031	0.716±0.022	0.727±0.023
NewsM-3-8-2	**0.596±0.039**	0.594±0.031	0.582±0.042	0.558±0.053	0.509±0.039	0.552±0.029	0.569±0.036	0.523±0.036
NewsNG-3-2-8	**0.898±0.019**	0.896±0.020	0.865±0.022	0.671±0.077	0.450±0.003	0.673±0.023	0.877±0.018	0.887±0.018
NewsNG-3-5-5	**0.751±0.025**	0.745±0.031	0.735±0.033	0.640±0.054	0.468±0.075	0.631±0.037	0.736±0.021	0.746±0.020
NewsNG-3-8-2	**0.616±0.033**	0.610±0.044	0.607±0.038	0.536±0.037	0.491±0.017	0.540±0.027	0.581±0.041	0.590±0.043
W / T / L	MuVAD-QPR vs. others		14 / 1 / 0	15 / 0 / 0	15 / 0 / 0	14 / 1 / 0	15 / 0 / 0	12 / 3 / 0
W / T / L	MuVAD-FSR vs. others		12 / 3 / 0	15 / 0 / 0	15 / 0 / 0	14 / 1 / 0	14 / 1 / 0	11 / 4 / 0



(a) QPR

(b) FSR

Figure 4: Illustration of instance weights distribution.



(a) QPR

(b) FSR

Figure 5: Convergence curves w.r.t. iteration time.

Table 3: AUC on CASME datasets.

	OCSVM	HOAD	CC	MLRA	DMOD	CRMOD	MuVAD
CAS-1	0.478	0.594	0.479	0.651	0.750	0.629	**0.815**
CAS-2	0.554	0.487	0.594	0.800	0.913	0.827	**0.953**
CAS-3	0.508	0.380	0.429	0.422	0.693	0.680	**0.725**
CAS-4	0.707	0.496	0.577	0.407	0.660	0.628	**0.801**
CAS-5	**0.675**	0.592	0.655	0.560	0.555	0.617	0.663
CAS-6	0.543	0.598	0.471	0.643	0.890	0.680	**0.896**
CAS-7	0.748	0.332	0.451	0.342	**0.986**	0.765	0.975
CAS-8	0.612	0.616	0.420	0.329	0.685	0.705	**0.713**
Mean	0.603	0.512	0.509	0.519	0.766	0.691	**0.818**

than MuVAD-QPR.

Micro-Expression Detection

This experiment evaluates the performance of the MuVAD on micro-expression detection task. We apply MuVAD-QPR on The Chinese Academy of Sciences Micro-Expression (CASME) dataset (Yan et al. 2013). CASME contains 195 micro-expressions filmed under 60fps. Each micro-expression is in a video clip (frame sequence), where the onset and offset frame (the first and the last frame of the micro-expression) are recorded. All frames of a video clip are organized as a multi-view dataset, where the first view is the upper half face and the second view is the lower half face. The frames from onset frame to offset frame are labeled as anomalies.

Experiments are performed on 8 frame sequences from CASME dataset. For each frame sequence, we eliminate the unnecessary background. The features of two views are extracted by the method of Local Binary Patterns (LBP) (Ahonen, Hadid, and Pietikäinen 2006). We set the diagram to 6 and the number of neighbors to 16. LBP histograms are then used as feature vectors. The experimental results are presented in Tab. 3. The proposed approach also gets best result on this task.

Conclusion

This paper proposes the MuVAD approaches focusing on addressing the main issues of existing multi-view anomaly detection approaches, which include clustering assumption on data, sub-optimal performance due to normal instance profiling and lack of robustness. Specifically, we propose a novel nearest neighbor-based anomaly measurement criterion firstly, then utilize this criterion to formulate an objective to estimate the set of normal instances explicitly. We develop two concrete relaxations for implementing the MUVAD as MUVAD-QPR and MUVAD-FSR. The MUVAD approaches are capable of handling the above mentioned issues of existing approaches. Experiments on datasets from different domains and real world application demonstrate the superiority of the proposed MUVAD approaches.

References

Ahonen, T.; Hadid, A.; and Pietikäinen, M. 2006. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 28(12):2037–2041.

Akoglu, L.; Tong, H.; and Koutra, D. 2015. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery* 29(3):626–688.

Aleskerov, E.; Freisleben, B.; and Rao, B. 1997. CARDWATCH: a neural network based database mining system for credit card fraud detection. In *Proceedings of IEEE Computational Intelligence for Financial Engineering*, 220–226.

Alvarez, A. M.; Yamada, M.; Kimura, A.; and Iwata, T. 2013. Clustering-based anomaly detection in multi-view data. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 1545–1548.

Bay, S. D., and Schwabacher, M. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 29–38.

Bickel, S., and Scheffer, T. 2004. Multi-view clustering. In *Proceedings of the 4th IEEE International Conference on Data Mining*, 19–26.

Bisson, G., and Grimal, C. 2012. Co-clustering of multi-view datasets: A parallelizable approach. In *Proceedings of the 12th IEEE International Conference on Data Mining*, 828–833.

Blum, A., and Mitchell, T. M. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 92–100.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM computing surveys* 41(3):15:1–15:58.

Ding, Q.; Katenka, N.; Barford, P.; Kolaczyk, E. D.; and Crovella, M. 2012. Intrusion as (anti)social communication: characterization and detection. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 886–894.

Duh, K.; man Au Yeung, C.; Iwata, T.; and Nagata, M. 2013. Managing information disparity in multilingual document collections. *ACM Transactions on Speech and Language Processing* 10(1):1:1–1:28.

Ekman, P. 2009. Lie catching and microexpressions. *The Philosophy of Deception* 118–133.

Gao, J.; Fan, W.; Turaga, D. S.; Parthasarathy, S.; and Han, J. 2011. A spectral framework for detecting inconsistency across multi-source object relationships. In *Proceedings of the 11th IEEE International Conference on Data Mining*, 1050–1055.

Jia, Y.; Salzmann, M.; and Darrell, T. 2010. Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems 23*, 982–990.

Kumar, A.; Rai, P.; and Daumé III, H. 2011. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems 24*, 1413–1421.

Li, S.; Shao, M.; and Fu, Y. 2015. Multi-view low-rank analysis for outlier detection. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, 748–756.

Liu, A., and Lam, D. N. 2012. Using consensus clustering for multi-view anomaly detection. In *Proceedings of the IEEE Symposium on Security and Privacy Workshops*, 117–124.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining*, 413–422.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* 13(7):1443–1471.

Spence, C.; Parra, L.; and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 3–10.

Wang, W., and Zhou, Z.-H. 2010. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, 1135–1142.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *CoRR* abs/1304.5634.

Yan, W.-J.; Wu, Q.; Liu, Y.-J.; Wang, S.; and Fu, X. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–7.

Ye, H.-J.; Zhan, D.-C.; Miao, Y.; Jiang, Y.; and Zhou, Z.-H. 2015. Rank consistency based multi-view learning: A privacy-preserving approach. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 991–1000.

Zhao, H., and Fu, Y. 2015. Dual-regularized multi-view outlier detection. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 4077–4083.

Zhao, H.; Liu, H.; Ding, Z.; and Fu, Y. 2018. Consensus regularized multi-view outlier detection. *IEEE Transactions on Image Processing* 27(1):236–248.