Graph-based Multimodal Fusion with metric learning for multimodal classification

Michalis Angelou, Vassilis Solachidis, Nicholas Vretos, Petros Daras

Centre for Research and Technology Hellas
Information Technologies Institute, Thessaloniki, Greece

Abstract

In this paper, a graph-based, supervised classification method for multimodal data is introduced. It can be applied on data of any type consisting of any number of modalities and can also be used for the classification of datasets with missing modalities. The proposed method maps the features extracted from every modality to a space where the intrinsic structure of the multimodal data is kept. In order to map the extracted features of the different modalities into the same space and, at the same time, maintain the feature distances between similar and dissimilar modality data instances, a metric learning method is used. The proposed method has been evaluated on NUS-Wide, NTU-RGBD and AV-Letters multimodal datasets and has shown competitive results with the state-of-the-art methods in the field, while is able to cope with datasets with missing modalities.

Keywords: Multimodal fusion, multimodal metric learning, multimodal classification, distance graphs

1. Introduction

With the wide growth of processing power and network speed, a simple event can be described using different types of multimedia. While in the past, news web pages would present an event with textual descriptions and photos, nowadays photo captions, keywords and videos are also used. At the same time, the wide use of capturing devices in combination

Email addresses: mich_angelou@iti.gr (Michalis Angelou), vsol@iti.gr (Vassilis Solachidis), vretos@iti.gr (Nicholas Vretos), daras@iti.gr (Petros Daras)

with the easiness of their usage, their mobility capabilities and their low cost, made it easy to capture an event from different users. Moreover, the ability of sharing these data with the community, made available not only an event description with a large number of captured data, but also its enrichment with textual and/or symbolic (emoticons and alike) description attributed by the users. Therefore, nowadays an event representation consists of a multitude of types of multimedia data, captured by different devices and user provided textual and/or symbolic descriptions. Thus, the need to develop methods that use and analyze this multimodal information has been emerged.

In this paper, a general classification framework is proposed that can be used on any kind of data which describe objects, actions, contexts or other informative cues (hereinafter all called objects) with the only constraint that an object is described by more than one modalities. Modalities, in this context, are considered the image, the video, the audio, the text or, in general, any data type that can be acquired by a device or added by the user to describe an event.

In multimodal classification the objective is to assign a predefined label (class) to an object that is described by more than one modalities. Lately, multimodal classification gained more and more attention from the research community for two reasons: Firstly, because nowadays more information for an object can be extracted from multiple sources. Secondly, the processing power growth enables such analysis that was forbidden up until recent years due to the induced complexity from the use of multiple modalities for classification.

There is a multitude of different approaches for multimodal classification. The two major frameworks that are used are: 1) classifying each modality separately and fuse the classification output to take a final decision (late fusion), and 2) by fusing multiple modalities and classify them as a single entity (early fusion) [1]. Recently, a vast amount of effort has been invested on multimodal fusion based on the biologically inspired idea that multimodal fusion resembles the way the human brain works by using many input sources (senses) to classify objects [2, 3]. The proposed method makes use of multimodal fusion since, even if each separate modality adds information to the object under examination, the join analysis of the different modalities can also provide additional information.

In this paper a supervised multimodal fusion model is presented that fuses data both in the feature level and the decision level, preserving information conveyed by one modality to the other. The objective is to map all the extracted modalities' features into the same feature space, where distance between them can be measured, and therefore can be used in a classification algorithm. The main assumption is that, when all modalities of an object are on the same feature space, they can be processed independently and when fused, to provide more information about it. This also allows at a next step, to use any classification method on the projected features, taking decisions in higher semantic level.

The proposed method employs a supervised graph-based method that takes into account similarities between different modalities of different training samples. In order to measure the similarity between samples arriving from different modalities, a metric is defined that is learned in a supervised way.

A common issue arising in multimodal setups is the fact that data from one or more modalities are missing at the testing phase. In the taxonomy presented in [4], it is shown that these cases form a distinct category of multimodal machine learning methods. In the proposed method, data can be considered as multiple single-modality objects. It is interesting that the proposed method is able to handle objects with missing modalities both in training and in testing phases.

In this work, a novel approach for multimodal fusion for classification is proposed. The method is generic and can be applied on datasets with any number and type of modalities. The relationship between the modalities is used on every step of the procedure, so that most of the information that exists among the modalities is exploited. The proposed method can be applied even if modalities are missing in a number of samples. In this case, there is no need of generating the missing modalities, since the available ones can be used both for training and testing. Additionally, samples with missing modalities can also be mapped to the feature space and thus be correctly classified.

2. Related Work

Multimodal fusion has been used in various tasks, such as action or expression recognition [5] - [7], image/video classification [8, 9], speech recognition [10, 11], person/object/context recognition [12] - [16], medical diagnosis [17] - [19], etc. The data that are usually used in these tasks are acquired from a large variety of sensors, such as bio-sensors (smart-meter sensors or blood-pressure devices, fingerprints) [20] - [22], other passive sensors (Kinect and other cameras, smart-phones) [23] - [27] or user created information, such as text or tags [28] - [30].

There is a plethora of methods applying multimodal fusion on different levels partitioned based on the stage at which the data fusion takes place (early fusion, late fusion, feature level fusion, decision level fusion). We refer the reader to [1] for a comprehensive analysis and presentation of different fusion methods. The methods therein are categorized according to the diversity and the type of the multimodal data to be fused. According to [3], there are various issues that usually come up in multimodal fusion such as data imperfections caused by sensors, data outliers, time-invariant and varying with time data, different data preprocessing, different data dimensions.

Regarding multimodal data analysis, in [12] a framework for multimodal content retrieval is presented that supports retrieval of rich media objects as unified sets of different modalities (image, audio, 3D, video and text), by efficiently combining all unimodal heterogeneous similarities to a global one according to an automatic weighting scheme. In [20] Kalimeri et al. present a multimodal framework using the fusion of electroencephalography (EEG) and electro-dermal activity (EDA) signals, for assessing the emotional and cognitive experience of blind and generally visually impaired people, when navigating in unfamiliar indoor environments. In [19] the authors use grey-scale video (EEG) as one modality and the optical flow between frames as the other modality. Then they use deep neural network (DNN) with convolutional neural network (CNN) and recurrent neural network (RNN) for the EEG classification. In [21] Wagh et al. fuse fingerprint and Iris images, in person identification for security purposes. In [23] Patwardhan et al. propose a method for detecting aggressive

actions and anger by fusing joint position, movement, body posture, head gesture, face and speech data that have been captured by a Kinect device. In [24] Cricri et al. use video, audio and sensorial data captured by mobile phones to classify sports.

Most of the research conducted thus far on multimodal fusion, either concatenates the features and applies a classification method (e.g. [20, 21] - feature level fusion), or uses the classification results for each modality to improve the result by aggregation (e.g. [23, 26] - decision level fusion). Many methods use also both levels of fusion (hybrid fusion). These methods do not exploit the mutual relationship between the different modalities, rather than the joint effect that they have on the final decision.

To the best of our knowledge, the only methods that use the relationship among the different modalities in order to classify the multimodal input are described in [31, 24] and [32]. Yet, in [31] there is the necessity of generating the missing modalities using information from the existing modalities in order to provide results. A similar approach is used in [32], where information theory based measures are used to generate missing modalities. Therein, the concatenated modalities are used to train a single representation using neural networks. In [24] a feature weight is calculated using the relation between different modalities through concatenation of the modalities and unimodal classification. Then, late fusion is applied in order to enhance the classification results using the previously calculated weights.

The proposed method builds on [33] and [29] which also deal with multimodal classification. In [33] a graph-based method is presented that performs multimodal image classification. However, this method has restrictions, since it cannot be applied on data consisting of modalities of different size or type. In fact, the method uses only one data type (image), while different poses of the same person are considered as different modalities. Graph-based classification has also been studied in [34]. This method gives good results in terms of image tagging, however it has not been used for multimodal classification purposes instead for feature learning and image understanding. In [29] a method that seeks a metric to be used as a distance between different modalities is proposed. This method maps all modalities to a single representation in a lower-dimensional space. However, this method cannot be applied on datasets that do not contain information for all modalities for each instance. Under the

same motivation, namely the need of measuring the similarity between different features, in [35] a method is proposed which uses a deep learning approach. This method also does not cope with multimodal classification since it is used solely for image retrieval. Our method employs matrix factorization in multiple steps of the process. Matrix factorization has also been studied for image understanding in [36], using a deep learning approach.

3. The Proposed Method

The general form of the proposed framework consists of six parts that are described in this section, and can be seen in Figure 2. Initially, the data are normalized by subtracting the mean value of each feature of each modality. Then, a transformation is applied using Singular Value Decomposition (SVD) so that all modalities lie on the same space, which is the space of one existing modality.

At a next step, the unimodal data-input is mapped into multiple representations in a new space, where distance between data can be measured. The representation in the new space is calculated using the right Truncated SVD transformation [37] matrix as a base. Truncated SVD is used to reduce the dimensions of the data with the minimum loss of information. Then the transformation matrix is updated so that the distance between data of the same/different class (label) is decreased/increased respectively [29]. The goal here is to find a distance metric between the different modalities. So the different modalities are mapped into a space where the distance between data of the same/different class (label) is small/large and hence use this mapping for calculating distances between modalities. Each update of the matrix results to a mapping that better complies to the general goal.

After this step, the different modalities (after being transformed) are used separately and passed to the graph-based method. In this phase, the graph-based method is used to preserve the distances between different objects. Then, the target representation in the new space is derived using the graph. On this final space, any single modality classification method can be applied.

Thus, the proposed method consists of the following steps:

- Normalization and mapping of the extracted features of all modalities to the feature space of one modality.
- Initialization of the first transformation matrix.
- Updating of the first transformation matrix based on the distance between modalities.
- Estimation of the final transformation matrix using a graph-based method.

3.1. Notations

Let $\mathbf{X}$ be a multimodal dataset of m modalities consisting of n labeled samples with label l_i, $i \in \{1, \ldots, c\}$. l_i represents the class of the i^{th} sample where c is the total number of the different classes. The representation of each sample of the modality $j \in \{1, 2, \ldots, m\}$ is given by the vector $\mathbf{x}_{i,j} \in R^{d_j}$ where d_j is the dimension of the sample representation for the modality j. Each sample can be represented by one vector $\mathbf{x}_i$ by concatenating its single-modality representations. Namely:

$$\mathbf{x}_i = [\mathbf{x}_{i,1}| \ldots |\mathbf{x}_{i,j}], j = 1, 2, \ldots, m \tag{1}$$

where $\mathbf{x}_i$ is of size $[1 \times d]$, $d = \sum_{j=1}^{m} d_j$. The entire data set is represented by the matrix $\mathbf{X}$ of size $n \times m$ where each line of $\mathbf{X}$ is the sample representation $\mathbf{x}_i$ namely:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \left[\begin{array}{c|c|c} \mathbf{x}_{1,1} & \ldots & \mathbf{x}_{1,m} \\ \vdots & & \vdots \\ \mathbf{x}_{n,1} & \ldots & \mathbf{x}_{n,m} \end{array}\right] = [\mathbf{X}_1| \ldots |\mathbf{X}_m] \tag{2}$$

In other words, the k^{th} value of the j^{th} modality of sample i is located at the i^{th} row and the $(f(j) + k)^{th}$ column of $\mathbf{X}$, where $f(j) = \sum_{q=1}^{j-1} d_j$. $\mathbf{X}_j$ denotes the $n \times d_j$ sub-matrix of $\mathbf{X}$ that contains the j^{th} modality features of all n samples.

3.2. Data pre-processing

The mean value of each modality per dimension is subtracted. For each sub-matrix $\mathbf{X}_j$ of the multimodal dataset, the mean value is calculated as:

$$\boldsymbol{\mu}_j = [\mu_{j,1}, \ldots, \mu_{j,d_j}] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{i,j} \tag{3}$$

and $\mathbf{z}_{i,j} = \mathbf{x}_{i,j} - \boldsymbol{\mu}_j$, with $\mathbf{Z}$ is $\mathbf{X}$ after normalization.

In the next step, all modalities $j, w \in \{1, 2, \ldots, m\}, j \neq w$ are transformed such that the distance $||\mathbf{Z}_w - \mathfrak{R}_j \mathbf{Z}_j||$ is minimized. The rotation matrices $\mathfrak{R}_j$ for each modality $j, w \in \{1, 2, \ldots, m\}$, $j \neq w$ are calculated.

All the modalities, except for the w^{th}, are transformed since for modality w, $\mathfrak{R}_w = \mathbf{I}_{d_w}$, where $\mathbf{I}_{d_w}$ is the $[d_w \times d_w]$ identity matrix.

The solution of this minimization problem is obtained using SVD over the covariance matrix $\text{cov}(\mathbf{Z}_w, \mathbf{Z}_j) = \mathbf{Z}_w^T \mathbf{Z}_j$.

$$\begin{aligned} \mathbf{Z}_w^T \mathbf{Z}_j &= \mathbf{U}_\mathfrak{R} \boldsymbol{\Sigma} \mathbf{V}_\mathfrak{R}^T \\ \mathfrak{R} &= \mathbf{V}_\mathfrak{R} \mathbf{I} \mathbf{U}_\mathfrak{R}^T \end{aligned} \tag{4}$$

where $\mathbf{V}_R$ is a matrix of size $[d_j \times d_j]$, $\mathbf{I}$ is a matrix of size $[d_j \times d_w]$ where its elements equal Kronecker delta ($I_{ij} = \delta_{ij}$), and $\mathbf{U}_\mathfrak{R}$ is a matrix of size $[d_w \times d_w]$.

Hereafter, $\mathbf{X}$ will indicate the transformed modalities with $\mathbf{X}_i = \mathfrak{R}_i \cdot \mathbf{Z}_i, \forall i \in \{1, 2, \ldots, m\}$ are then concatenated as in (2) resulting in matrix $\mathbf{X}$.

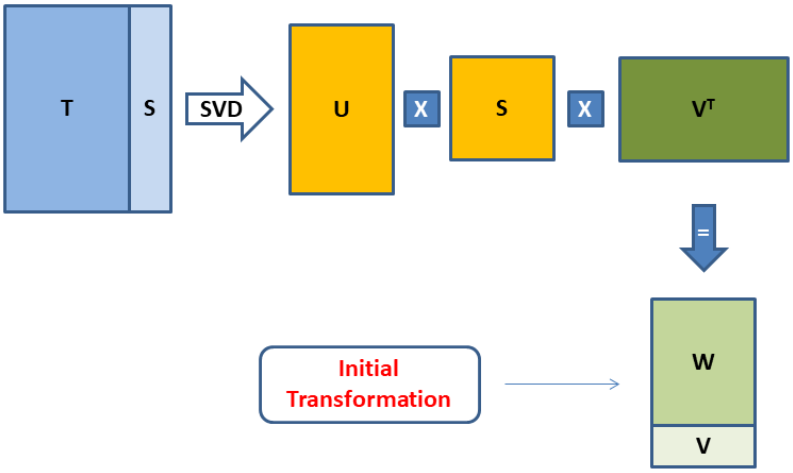
3.3. Initial transformation

Truncated SVD Decomposition is applied on the above dataset for s largest singular values so that $\mathbf{X} \approx \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T$. $\mathbf{V}$ is the initial transformation matrix of size $[d \times s]$ where s is manually selected and since $s < d$ dimensionality reduction of the dataset is also achieved. $\mathbf{V}$ is the initial transformation matrix. It can be further updated by taking into account the class information.

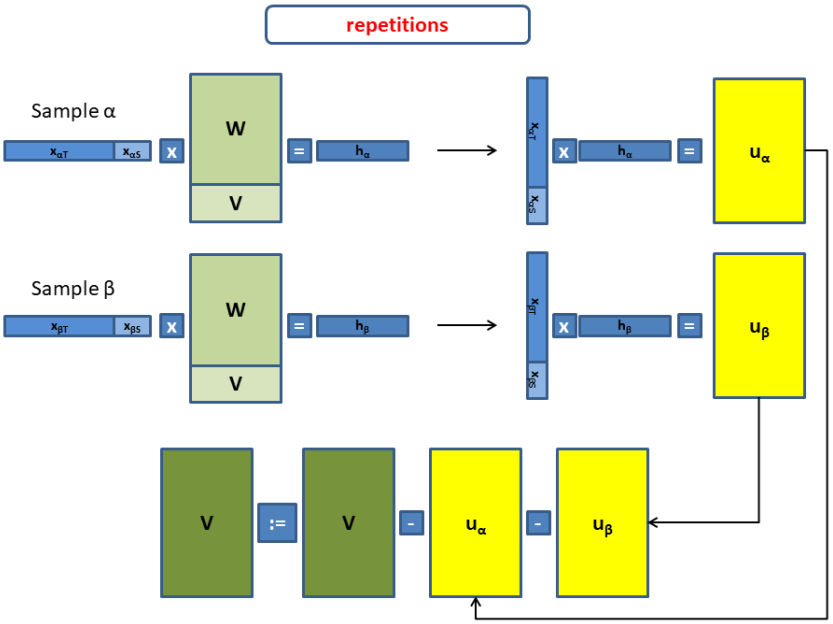
An iterative update algorithm is applied on the transformation matrix $\mathbf{V}$ (Algorithm 1). Considering $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$ to be two random samples, for all pairs $\mathbf{x}_\alpha$, $\mathbf{x}_\beta$, we compute:

$$\mathbf{h}_\alpha = \mathbf{x}_\alpha \cdot \mathbf{V} \text{ and } \mathbf{h}_\beta = \mathbf{x}_\beta \cdot \mathbf{V} \tag{5}$$

where $\mathbf{h}_\alpha$ and $\mathbf{h}_\beta$ are the representations of the two respective samples and $\alpha, \beta \in \{1, 2, \ldots, n\}, \alpha \neq \beta$.



(a) The SVD that provides the Initial transformation matrix $\mathbf{V}$.



(b) The repetitive procedure that updates the initial matrix to the final

Figure 1: The transformation of data for the initial transformation matrix $\mathbf{V}$. $\mathbf{T}$ corresponds to the first while $\mathbf{S}$ to the second modality. This figure refers to the realization on NUS-Wide 1.5K data.

Then, the updated values for the transformation matrix $\mathbf{V}$ are calculated by:

$$\begin{aligned} \mathbf{u}_\alpha &= \mathbf{x}_\alpha^T \cdot (\mathbf{h}_\alpha - \mathbf{h}_\beta) \cdot 2 \cdot C_1 \cdot st \\ \text{and } \mathbf{u}_\beta &= \mathbf{x}_\beta^T \cdot (\mathbf{h}_\beta - \mathbf{h}_\alpha) \cdot 2 \cdot C_2 \cdot st \end{aligned} \tag{6}$$

where C_1 and C_2 are parameters balancing the significance between the update matrices $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$, and st is the step-size parameter that determines how quickly the transformation matrix will converge to the final matrix.

The transformation matrix $\mathbf{V}$ is updated as shown below for the pairs of α and β belonging to the same cluster ($l_\alpha = l_\beta$):

$$\mathbf{V} := \mathbf{V} - (\mathbf{u}_\alpha + \mathbf{u}_\beta) \tag{7}$$

while for the pairs of α and β with $l_\alpha \neq l_\beta$, $\mathbf{V}$ is updated iff the euclidean distance between $\mathbf{h}_\alpha$ and $\mathbf{h}_\beta$ is below a threshold ϵ:

$$\mathbf{V} := \mathbf{V} + (\mathbf{u}_\alpha + \mathbf{u}_\beta) \text{ , iff } \|\mathbf{h}_\alpha - \mathbf{h}_\beta\|_2 \leq \epsilon \tag{8}$$

This process is repeated for r iterations, so that the $\mathbf{V}$ matrix best represents the similarities between modalities, after a predefined amount of iterations. The entire process for the bi-modal case ($m = 2$) is illustrated in Figure 1.

3.4. Reshaping the dataset

After r iterations, the resulted transformation matrix is $\mathbf{V}$. $\mathbf{V}$ consists of m concatenated matrices $\mathbf{V}_j$

$$\mathbf{V} = \Big[\mathbf{V}_1 \cdots \mathbf{V}_m\Big]^T \tag{9}$$

where each $\mathbf{V}_j$ matrix represents the j^{th} modality of the dataset and is of size $[d_j \times s]$, while s is the number of the largest singular values of the truncated SVD.

In the next step, the dataset is divided into m parts, one for each modality. The reshaped dataset can be written as :

$$\mathbf{X}' = \Big[\mathbf{X}'_1 \cdots \mathbf{X}'_m\Big] \tag{10}$$

Algorithm 1 The SGD Updating Algorithm

Initialize $\mathbf{V} = \mathbf{V}_0$ by applying SVD

repeat

  repeat

    given a pair of samples (α, β) compute:

    $\mathbf{h}_\alpha$ and $\mathbf{h}_\beta$

    $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$

    if α and β similar $(l_\alpha = l_\beta)$ **then**

      $\mathbf{V} := \mathbf{V} - (\mathbf{u}_\alpha + \mathbf{u}_\beta)$

    else

      if $\|\mathbf{h}_\alpha - \mathbf{h}_\beta\|_2 \leq \epsilon$ **then**

        $\mathbf{V} := \mathbf{V} + (\mathbf{u}_\alpha + \mathbf{u}_\beta)$

      end if

    end if

  until a predefined number of pairs

until predefined number of iterations

where $\mathbf{X'}_j = \mathbf{X}_j \cdot \mathbf{V}_j$ is a sub-matrix of $\mathbf{X'}$ that contains the transformation of the j^{th} modality of all samples, the size of $\mathbf{X'}_j$ is $[n \times s]$, hence the size of $\mathbf{X'}$ is $[n' \times s]$, where $n' = n \cdot m$.

The above derives from the fact that by using simple matrix multiplication we can take:

$$\mathbf{X} \cdot \mathbf{V} = \mathbf{X}_1 \cdot \mathbf{V}_1 + \cdots + \mathbf{X}_m \cdot \mathbf{V}_m = \sum_{j=1}^{m} \mathbf{X}_j \cdot \mathbf{V}_j \tag{11}$$

This is the point where we intervene in the original method that is proposed in [29], where a single representation for the concatenation of the two modalities is calculated. At this representation a distance between the different objects consisting of all modalities can be measured. We assume that we can keep the parts of the $\mathbf{V}$ that correspond to the j^{th} modality in order to map each modality to the new feature space independently. In this new space, the distance can also be measured between different modalities.

3.5. Graphs Creation

For the reshaped dataset $\mathbf{X}'$, two graphs are constructed. A between-class graph $\{G_b, \mathbf{W}_b\}$ and a within-class graph $\{G_w, \mathbf{W}_w\}$, with $\mathbf{W}_b$ and $\mathbf{W}_w$ being the weight matrices of the two graphs, respectively.

To make the connections between the different nodes of the graphs and thus create meaningful edges, we use the kernel-based distances that we consider as similarity of the nodes. To do so, we have made the hypothesis that the data lie on multiple manifolds in order to put constraints for the connections between the nodes, as will be described in the following paragraphs. Moreover, the evaluation of the kernel-based distances will be used as weights of the edges of the graphs.

Given n' data samples $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}]^T$ from c different classes, the data-points are separated into $\varrho = c \cdot m$ modality manifolds $M = \{M_1, \ldots M_\varrho\}$, where $M_{j,l}$ is defined as the j^{th} modality-manifold-fragment (MMF) of the l^{th} class. Then, the within-class graph $\{G_w, \mathbf{W}_w\}$ and the between-class graph $\{G_b, \mathbf{W}_b\}$ graphs are constructed.

1) Within-class graph $\{G_w, \mathbf{W}_w\}$. Within-class graph is based on the MMF inner structure. Thus, $\forall \mathbf{x}_i \in M_{l,k}$ $(l = 1, \ldots, \rho,\ k = 1, \ldots, c)$ we connect $\mathbf{x}_i$ with all $\mathbf{x}_j \in Q \subset M^i_{l,k}$ where $|Q| = k$, $d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{x}_k)$, $\forall j \in Q$, $k \in M^i_{l,k} - Q$ where $M^i_{l,k} \subset M_{l,k}$ with all the elements of $M_{l,k}$ excluding the ones that have been connected with $\mathbf{x}_i$ in previous steps and $\mathbf{x}_i$ itself.

In other words, for all vertices that belong to the same MMF, namely the samples that belong to the same class and the same modality, $\mathbf{x}'_{i'} \in M^M_{j,l}$, an edge is added between $\mathbf{x}'_{i'p}$ and $\mathbf{x}'_{i'q}$, if $\mathbf{x}'_{i'q}$ is among the k nearest-neighbors of $\mathbf{x}'_{i'p}$. In the case that an edge already exists from a previous iteration of the algorithm, the next nearest-neighbor is selected.

The process above is performed for every MMF. In the next step, edges are added between different MMFs that correspond to the same class using the following procedure:
$\forall$ MMF $M_{i,k}$, connect $M_{i,k}$ with $M_{j,k}$ if $\mathbf{x}_a \in M_{j,k} : d(\mathbf{x}_a, \mathbf{x}_b) < d(\mathbf{x}_c, \mathbf{x}_d)$ $\forall \mathbf{x}_a, \mathbf{x}_c \in M_{i,k}$ and $\mathbf{x}_b, \mathbf{x}_d \in M_{l,k}$.

Thus, an edge is added between the two closest vertices belonging to different MMFs, with the same restrictions set for the inner MMF structure. Yet, two MMFs are considered

connected if any two of their vertices are connected.

2) Between-class graph $\{G_b, \mathbf{W}_b\}$. The MMF inner connections are used as described above. Then, edges are added between different MMFs that correspond to the same modality. The graph vertices are connected with the same restrictions as set in the within-class graph construction.

3.6. Graph output

From the graphs created in the previous step, the corresponding adjacency matrices $\mathbf{A}_w$ and $\mathbf{A}_b$ are obtained. The $\mathbf{W}_w$ and $\mathbf{W}_b$ weight matrices are calculated using the heat kernel according to which the weights of an edge between two vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ are given by:

$$\mathbf{W}_{i,j} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{h}} \tag{12}$$

The two laplacian matrices $\mathbf{L}_w$ and $\mathbf{L}_b$ are then calculated as follows:

$$\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w \text{ and } \mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b \tag{13}$$

where $\mathbf{D}_w, \mathbf{D}_b$ are the diagonal matrices with the row-sums of the weight matrix as the elements of the main diagonal, namely $\mathbf{D}_{w_{i,i}} = \sum_j \mathbf{W}_{w_{i,j}}$.

3.7. Final data representation calculation

As described in detail in [33], the objective is to simultaneously minimize three different quantities:

1. $\mathbf{Y}^T\mathbf{L}_w\mathbf{Y}$ s.t $\mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = \mathbf{I}$
2. $\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2$
3. $\|\mathbf{A}\|_{2,1}$

The final objective function is the following minimization of the weighted sum of the three quantities above, namely:

$$\min(\mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1}) \text{ s.t. } \mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = I \tag{14}$$

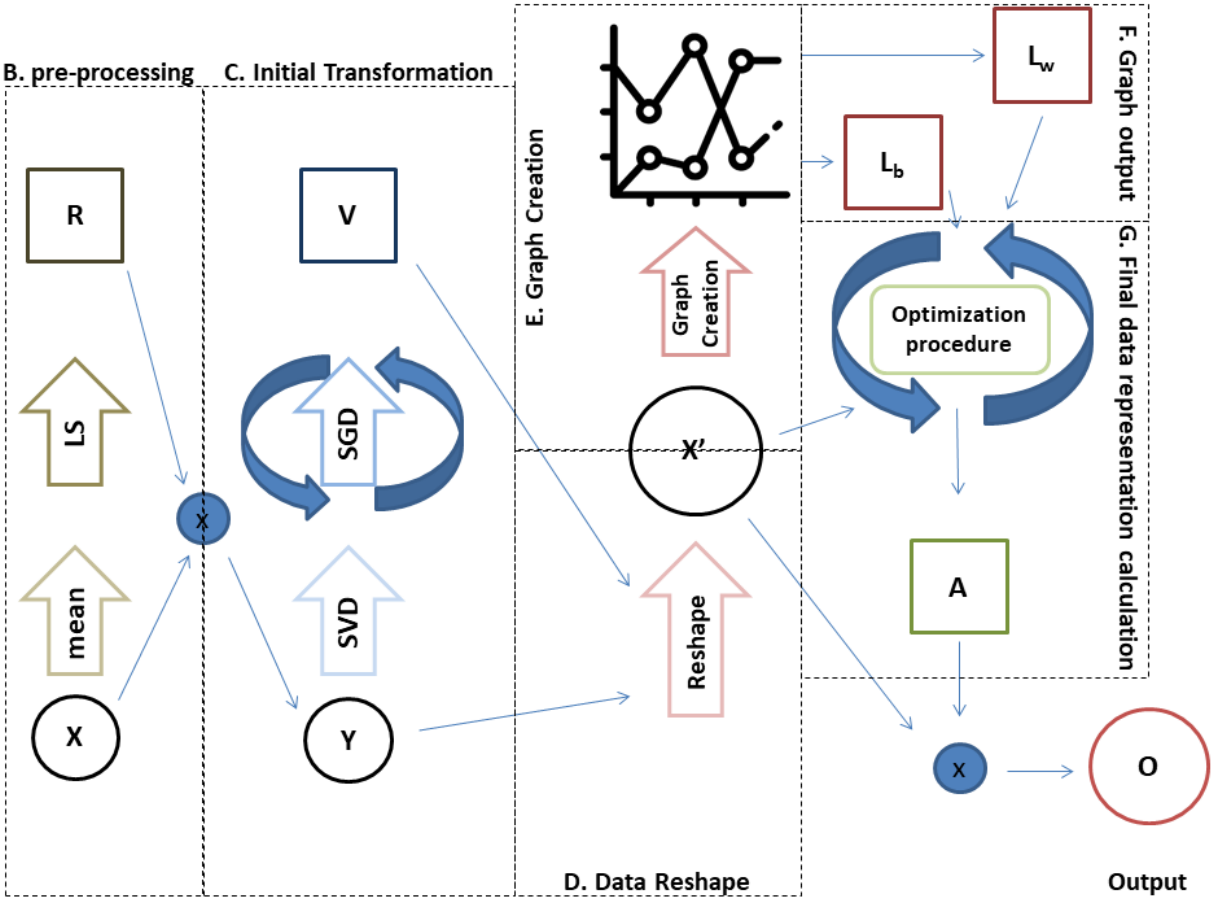


Figure 2: The different steps of the proposed method as they are described in 3. The titles of the parts are following the corresponding subsections in which the parts are described.

where ϖ and σ are two balance parameters, $\mathbf{A}$ is a transformation matrix, and $\|A\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{p}([\mathbf{A}]_{ij})^2}$ or zero otherwise. By setting the objective function as $\mathcal{F}$ we get:

$$\mathcal{F} = \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1} \tag{15}$$

By differentiating $\mathcal{F}$ with respect to $\mathbf{A}$, setting it to zero and solving for $\mathbf{A}$, we get:

$$\mathbf{A} = (\mathbf{X}\mathbf{X}^T + \sigma\mathbf{\Delta}/2\varpi)^{-1}\mathbf{X}\mathbf{Y} = \hat{\mathbf{A}}\mathbf{Y} \tag{16}$$

where $\mathbf{\Delta}$ is a diagonal matrix whose i^{th} diagonal element $\Delta_{i,i}$ equals to $(\|\boldsymbol{\alpha}_i\|)^{-1}$ only when $\boldsymbol{\alpha}_i \neq 0$ ($\boldsymbol{\alpha}_i$ is the i^{th} row vector of $\mathbf{A}$). The proof of (16) is given in Appendix A. Here we result in a different equation than the corresponding one in [33].

$$
\begin{aligned}
\mathcal{F} &= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1}) \\
&= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A} - 2\mathbf{A}^T\mathbf{X}\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \sigma\mathbf{A}^T\mathbf{\Delta}\mathbf{A} \\
&= \mathbf{Y}^T\mathbf{L}_w\mathbf{Y} + \varpi(\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}}\mathbf{Y} - 2\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{X}\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \\
&\qquad + \sigma\mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}\mathbf{Y} \\
&= \mathbf{Y}^T[\mathbf{L}_w + \varpi(\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}} - 2\hat{\mathbf{A}}^T\mathbf{X} + \mathbf{I}) + \sigma\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}]\mathbf{Y}
\end{aligned} \tag{17}
$$

If we set: $\mathcal{L} = \mathbf{L}_w + \varpi(\hat{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\hat{\mathbf{A}} - 2\hat{\mathbf{A}}^T\mathbf{X} + \mathbf{I}) + \sigma\hat{\mathbf{A}}^T\mathbf{\Delta}\hat{\mathbf{A}}$ the minimization function can be re-written as:

$$
\min(\mathbf{Y}^T\mathcal{L}\mathbf{Y}) \text{ s.t. } \mathbf{Y}^T\mathbf{L}_b\mathbf{Y} = \mathbf{I} \tag{18}
$$

According to the Lagrangian method, the optimization problem can be solved by computing the eigenvectors corresponding to the l smallest eigenvalues of the following generalized eigenvector problem:

$$
\mathcal{L}\mathbf{Y} = \lambda\mathbf{L}_b\mathbf{Y} \tag{19}
$$

The optimization process is illustrated in Algorithm 2:

Algorithm 2 Optimization Algorithm

Initialize $\mathbf{\Delta_0} = \mathbf{I}$, $t = 0$, ϖ, σ

repeat

compute $\mathbf{Y}_t$ by solving $\mathcal{L}\mathbf{Y} = \lambda\mathbf{L}_b\mathbf{Y}$

update $\mathbf{A}_t$ using $\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T + \frac{\sigma\mathbf{\Delta}}{2\varpi}\right)^{-1}\mathbf{X}\mathbf{Y}$

evaluate $\mathbf{\Delta}$ from $\mathbf{A}_t$

$t = t + 1$

until convergence or preset iterations

Hence, the final representation of the (pre-processed) input data is

$$
\mathbf{R} = \mathbf{x} \cdot \mathbf{V} \cdot \mathbf{A} \tag{20}
$$

which can then be passed into any single modality classification method.

In the testing procedure, let $\mathbf{O}$ be an object to be classified that consists of $z \leq m$ modalities. Let also $\mathcal{Z} = \{z_1, \ldots, z_z\} \subset \{1, \ldots, m\}$ where $\mathcal{Z}$ is the set of the indices of the existing modalities of the object $\mathbf{O}$ and m the number of modalities of the training set. Then, $\mathbf{O}$ is represented as

$$\mathbf{O} = \frac{\sum_{i=1}^{z}(\mathbf{x}_{O,i} \cdot \mathbf{V}_{z_z} \cdot \mathbf{A})}{z} \tag{21}$$

where, $\mathbf{x}_{O,i}$ is the representation of the i^{th} modality of $\mathbf{O}$, $\mathbf{V}_{z_z}$ is the ${z_z}^{th}$ submatrix of $\mathbf{V}$ (see (9)). It should be noted that similarly to the training procedure, during testing the input data are normalized and rotated by $\mathfrak{R}$. Thus, the $\boldsymbol{\mu}_j$ vectors from the training step are kept and re-used during the testing.

At this point we want to note that in case where there are samples in the training set with missing modalities, they can be included in the training procedure, skipping though some steps. If a modality of a sample in training is missing, the specific sample does not take part in the pre-processing modality-transformation part and the metric learning part (in initialization only). Yet in the next step, the existing modalities (except the w^{th}) will be transformed using (4). The only pre-processing the specific sample is through is the normalization.

4. Experimental Results

Experiments were conducted on three multimodal datasets: NUS-Wide [38], NTU RGB-D [39] and AV-Letters [40].

NUS-Wide samples consist of two modalities, images (six types of low-level features extracted from them) and their associated tags from Flickr. We used a subset of 1520 samples from the dataset (NUS-Wide 1.5K) and kept the bag-of-words feature based on SIFT descriptors.

NTU RGB-D dataset contains samples acquired by Microsoft Kinect devices. Each sample represents a sequence of frames. In each frame, one or two bodies are recorded. For each body, the 25 skeleton joints provided by the Kinect, are stored. For every joint, x, y, z (3D) coordinates of the joint, 2D (x, y) mapping of the corresponding depth frame,

2D (x, y) mapping of the corresponding RGB frame and 4D (w,x,y,z) orientation of the joint are provided. From the entire dataset, 500 samples are selected randomly to work on.

The AV-Letters dataset consists of 780 samples. Each of the 780 samples contains a sequence of video frames of various length illustrating lip-movement pronouncing a letter (video) and a sequence of Mel Frequency Cepstral Coefficients (MFCC) describing the audio of the corresponding letter.

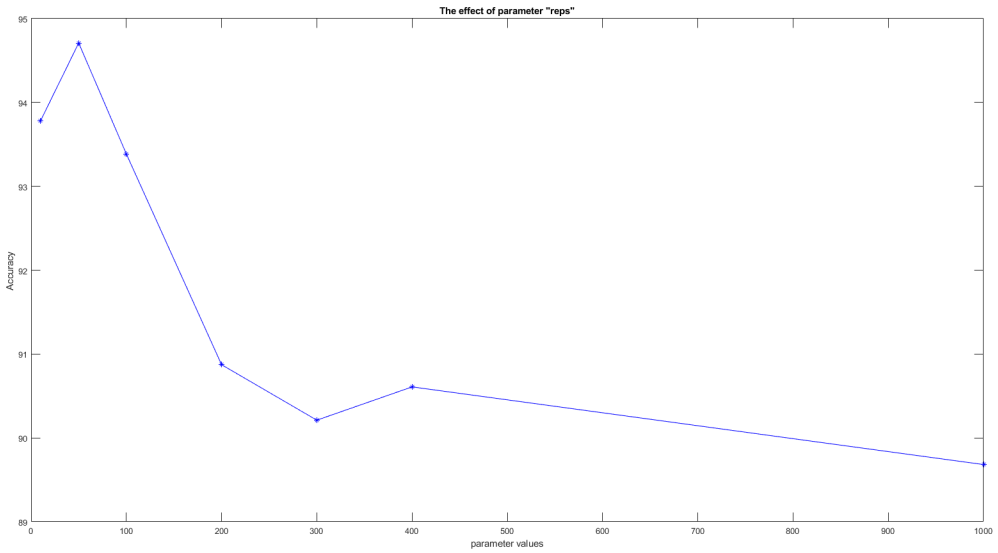
In order to estimate appropriate parameters for our method, an exhaustive heuristic method has been employed. The search concerned the ϖ, σ, t parameters, the number of singular values kept for the SVD step (sv), the number of iterations of the updating algorithm (r), the number of eigenvalues kept in the Graph Optimization algorithm (l) and the trade-off parameters (C_1, C_2), as described in Section 3. In Table 1, different indicative values of the parameters used in experiments are shown. The optimal values have been achieved for the parameters are written in bold. It is interesting that for most parameters the optimal values are close for the datasets described above. This fact indicates that the proposed method achieves strong generalization across datasets. The method's accuracy of multimodal classification on NUS-Wide 1.5k dataset for different values of parameters r and h is illustrated in Figures 3a and 3b respectively.

4.1. NUS-Wide 1.5K dataset

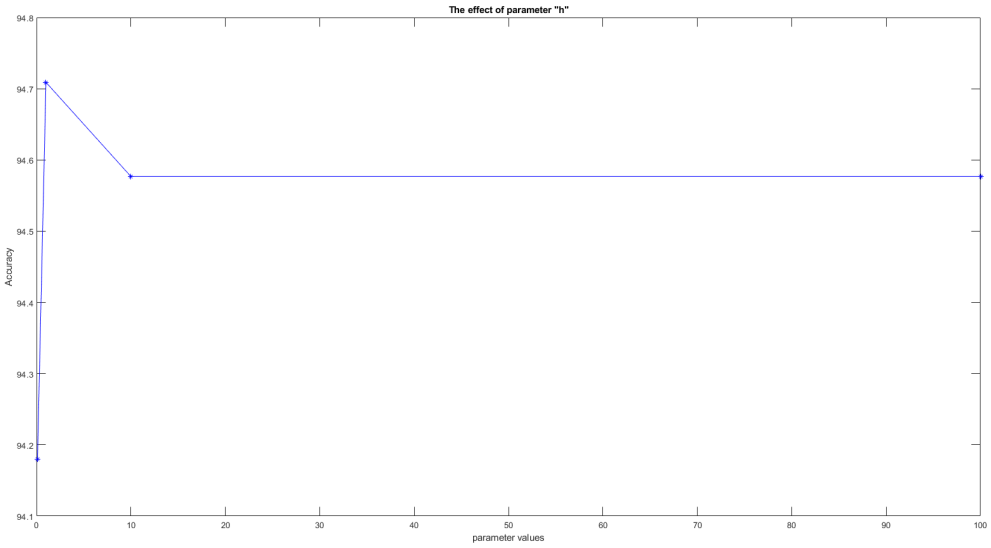
As mentioned in the previous subsection, each of the 1520 samples of the NUS-Wide 1.5K dataset consists of two modalities that describe an image. The first modality is a $1000d$ binary vector indicating the existence of 1000 tags on the image. The second modality is the probability of SIFT features that has been clustered into a bag-of-words of 1024 bins, to be found in the certain sample. A subset of 765 samples are used for training and a subset of 756 samples are used for testing.

The two modalities are concatenated following the procedure described in section 3.1 resulting in a vector $\mathbf{x}_i$ of length $d = 2024$. Then, the SGD step is applied on $\mathbf{V}$, for all the pairs of similar samples (9613 pairs), and for 10067 random dissimilar pairs.

In the next step, the dataset is reshaped (Section 3.4). Each of the new samples has



(a) The effect of r parameter on multimodal accuracy for the NUS-Wide 1.5k dataset.



(b) The effect of h parameter on multimodal accuracy for the NUS-Wide 1.5k dataset.

Figure 3: Some indicative illustration of the effect of the parameters on multimodal accuracy

length $sv = 60$ and is labeled with the class and the modality it belongs to. The optimization process is repeated until convergence or until 200 iterations.

An 8-fold cross validation is applied using multiple classifiers and the accuracy of each of different cases is computed. The classifiers are trained with the single representation (the mean of the resulting representations of each modality) and with independent representations

Table 1: Parameters used in the experiments

	NUS	**NTU**	**AV**
ϖ	**1**, 10, 100, 1000	1, **10**, 100, 1000	**1**, 10, 100
σ	1, **10**, 100, 1000	1, **10**, 100, 1000	**1**, 10, 100
h	0.1, **1**, 10, 100	0.1, **1**, 10, 100	0.1, **1**, 10, 100
sv	10, 30, 50, **60**, 100	10, **30**, 50, 60, 100	10, 30, **50**, 60, 100
l	10, 30, 50, **60**, 100	10, **30**, 50, 60, 100	10, 30, **50**, 60, 100
r	50, **100**, 200, 400	50, **100**, 200, 400	**50**, 100, 200
C_1	**100**, 200, 300, 400	**100**, 200, 300, 400	100, 200, **300**, 400
C_2	200, 300, **400**, 500	**100**, 200, 300, 400	200, **300**, 400, 500

Table 2: The experimental implementations. The enumeration of cases are as described in section 4.4.

	Framework Parts				Classifiers Training	
	LS	**SVD**	**SGD**	**Graph**	**Modalities sep**	**Modalities av**
1		X		X	X	
2		X		X		X
3		X	X	X	X	
4		X	X	X		X
5	X	X		X	X	
6	X	X		X		X
7	X	X	X	X	**X**	
8	X	X	X	X		X

Table 3: The experimental implementations with tested accuracy and best classifiers and the experimental results per modality for NUS-Wide. The enumeration of cases are as described in section 4.4.

	Multimodal CVA Accuracy		Per modality CVA Accuracy	
	av.	sep	Tags Mod	SIFT Mod
1	91.71%	48.98%	92.88%	5.13%
2	91.95%	48.82%	92.61%	4.28%
3	94.34%	50.53%	94.08%	6.97%
4	94.46%	49.70%	**94.60%**	4.87%
5	92.09%	49.48%	93.53%	5.42%
6	92.09%	49.60%	92.86%	5.81%
7	**94.74%**	**51.48%**	94.47%	**8.44%**
8	94.20%	49.64%	94.47%	4.87%

of each modality.

The complete parameter selection is shown in Table 1. The threshold for updating the $\mathbf{V}$ transformation matrix in case of dissimilar samples, is set to $\epsilon = 1$. The results on NUS-Wide dataset are presented in Table 2. The eight sub-cases (lines) of Table 3 correspond to experiments that are presented in detail in sub-section 4.4.

4.2. NTU RGB-D dataset

Each frame sequence of the NTU RGB-D dataset is considered as a sample and the Bag of Words (BoW) method is applied on the data. We use a random 85% of the selected samples for training and the rest for testing. All the 25×3, 25×2, 25×2 and 25×4 features of all frames are concatenated into 4 matrices and then k-means is applied on them for 50

words (centers) each. Then, the probability of each word appearing in the each sample is calculated. This procedure results in 4 sparse features of size (1×50), one for each sequence. Since each sample is constructed as the concatenation of the 4 modalities feature vectors, its dimension is (1×200). The SGD step is applied on $\mathbf{V}$ for the same number of pairs of similar/dissimilar samples (1423/1423 pairs).

For the next step the dataset is reshaped at $n' = n \cdot m = 4 \cdot 425 = 1700$ samples. The new sample length after SVD is $sv = 30$. The results are presented in Table 4. The $\mathcal{A}$ and $\mathcal{B}$ rows in the table show the baseline. The baseline for comparison purposes is on the extracted features of the BoW, without any processing. The same classifiers and the same classification procedure is applied on the features to show the difference in terms of accuracy.

Considering this dataset, results for one missing modality are presented in Table 5, and they are compared to the corresponding results of Table 4 under the *av.* column.

Table 4: The experimental implementations for NTU.

	MM CVA Accuracy		Per modality CVA Accuracy			
	av.	sep	Mod1	Mod2	Mod3	Mod4
$\mathcal{A}$	3.64%	57.45%	56.72%	59.15%	56.52%	57.35%
$\mathcal{B}$	61.54%	3.14%	3.23%	3.33%	5.00%	1.69%
1	50.00%	**62.50%**	**67.21%**	60.94%	62.50%	59.70%
2	60.00%	18.26%	11.86%	26.79%	32.79%	0.00%
3	55.00%	61.81%	60.61%	63.93%	**64.41%**	58.82%
4	58.21%	20.76%	18.75%	31.75%	24.14%	5.88%
5	3.77%	62.26%	60.94%	62.90%	60.61%	**64.62%**
6	57.58%	5.16%	2.08%	5.66%	5.88%	6.56%
7	40.98%	**62.50%**	60.00%	**65.00%**	64.06%	61.19%
8	**70.69%**	16.90%	18.97%	22.58%	19.61%	2.38%

Table 5: The experimental implementations for NTU for 1 missing modality.

	Overall output	Per missing modality CVA Accuracy			
		Mod1	Mod2	Mod3	Mod4
$\mathcal{B}$	61.54%	31.67%	31.58%	42.19%	28.07%
2	60.00%	40.91%	53.23%	50.72%	**50.77%**
4	58.21%	49.18%	50.00%	48.48%	47.69%
6	57.58%	9.43%	14.81%	17.46%	6.45%
8	**70.69%**	**51.61%**	**64.41%**	**60.32%**	43.94%

4.3. AV-Letters dataset

A neural network is used for feature extraction. The network consists of two LSTM layers. The two modalities are passed through two similar neural networks separately. The outputs of the networks are feature vectors for each modality of each sample, one for audio and one for video. The length of each feature vector is $[1 \times 26]$ and represents the possibility of the sequence to be classified in one of the 26 classes. We used a random subset of 650 samples for training and the rest 130 for testing. The features of the two modalities are then concatenated resulting in a vector $\mathbf{x}_i$ of length $d = 52$. The SGD step is applied on $\mathbf{V}$ for 8160 pairs of similar samples, and for 8160 random number of pairs of dissimilar samples.

Then, the dataset is reshaped at $n' = n \cdot m = 2 \cdot 650 = 1300$ samples resulting (after SVD) in sample length equal to $sv = 60$. Finally, as in NUS-Wide dataset, a 5-fold cross validation has been applied. The results for AV-Letters dataset are presented in Table 6.

4.4. Results interpretation

LS and SGD are procedures that are included/omitted in the experiments resulting in variations of the proposed framework as shown in the *Framework Parts* columns of Table 2.

As shown in (20), the training set elements that are fed into the classification method are the rows of $\mathbf{R}$. More specifically, we calculate the representation for each modality $\mathbf{R}_j$

Table 6: The experimental implementations for AV-Letters.

	MM CVA Accuracy		Per Modality CVA Accuracy	
	av	sep	mod-A	mod-V
1	70.77%	46.54%	33.21%	**59.87%**
2	72.82%	41.60%	23.97%	59.23%
3	70.77%	46.67%	33.46%	**59.87%**
4	72.69%	41.41%	23.33%	59.49%
5	71.15%	**47.05%**	34.36%	59.74%
6	**73.21%**	41.15%	22.95%	59.36%
7	71.15%	**47.05%**	**34.49%**	59.62%
8	73.08%	41.03%	22.56%	59.49%

namely

$$\mathbf{R}_j = \mathbf{X}_j \cdot \mathbf{V}_j \cdot \mathbf{A} \tag{22}$$

where $j = 1, 2, \ldots, m$. Each element $\mathbf{r}_{i,j}$

$$\mathbf{R} = \left[\begin{array}{c|c|c} \mathbf{R}_1 & \ldots & \mathbf{R}_m \end{array} \right] = \left[\begin{array}{c|c|c} \mathbf{r}_{1,1} & \ldots & \mathbf{r}_{1,m} \\ \vdots & & \vdots \\ \mathbf{r}_{n,1} & \ldots & \mathbf{r}_{n,m} \end{array} \right] \tag{23}$$

is the representation of the j^{th} modality of the i^{th} sample. Since all m modalities lie on the same space in the final representation, in the training procedure we can use from each sample either all the m l-d vectors (*Separate training*) or their average vector $R_i^{av} = \sum_{i=1}^{m} r_{i,j}$ (*Average Training*). Hence in the first case the number of training vectors per sample is m (in total $m \cdot n$ elements - **Modalities sep** column) while in the latter is one (in total n elements - **Modalities av** column), as indicated in Table 2 under *Classifier Training* columns.

Similarly, in the testing procedure the input consists of m l-d vectors for each sample. Thus, for the classification we can use the average of these m vectors (*Multimodal CVA accuracy - sum* column of Tables 2-6) or each modality representation separately (*Per modality CVA accuracy* columns), where CVA stands for cross validation average. In other words, on these columns, the method's accuracy is presented in the cases that all or only one modality is available. Column (*Multimodal CVA accuracy - sep* of Tables 2-6) equals the average of the m columns under *Per modality CVA accuracy*.

From Table 2 it seems that in the NUS-Wide dataset, *Tag* modality significantly outperforms *SIFT* modality, where the latter achieves very low accuracy ($< 10\%$). We also observe the same behavior in the AV-Letters dataset (Table 6) where *Mod-V* outperforms *Mod-A*.

During the data-preprocessing (Subsection 3.2), for the experiments on NUS-Wide and AV-Letters datasets, the modalities have been transformed to the first modality space ($w = 1$) as shown in Tables 2-6. This covers both possible cases, namely, in the NUS-Wide dataset we mapped the modality with the low accuracy to the one with the high accuracy, while in the AV-Letters we performed the inverse. We have also performed mappings to the other modalities ($w \neq 1$) which resulted in similar with the initial case ($w = 1$) accuracies. On the contrary, the results on NTU-RGBD dataset are for modalities transformed to the 4^{th} modality as shown in Table 4. Even though the 4^{th} modality shows the lowest accuracy on its own, when transforming all modalities to its space, the method shows significant improvement compared to transforming to the other modalities.

As expected, the results were better when the same training/testing object vector representation was used. Thus, the average of the modalities achieved better results in the classifier trained with the average of the modalities and vice versa.

The proposed method gives comparable results to the state-of-the-art methods. As can be seen in Tables 7 and 8, in AV-Letters database, our method surpasses the others by far for the multimodal case, even though for the case of the single-modal classification of AV-Letters is ranked last. In NUS-Wide database, our method surpasses previous state-of-the-art methods. In all cases the reader is referred to the referenced works for more

Table 7: Method performance in AV-Letters

	AV	**A**	**V**
MDAE [2]	62.90%	58.40%	62.10%
CRBM [31]	64.8%	61.2%	62.60%
RTMRBM [41]	66.04%	64.41%	64.63%
Proposed	**73.21%**	**34.49%**	**59.87%**

Table 8: Method performance in NUS-Wide 1.5K. The un-cited methods' results are taken from [29]

Xie [29]	93.52%	Xing+Original	89.95%
ITML+Original	89.95%	Xing+MWH	89.95%
ITML+MWH	92.86%	MKE	80.56%
Proposed	**94.74%**		

results compared to other methods. The numerical results given in the Tables for the other methods, are taken by [41] and [29], respectively. For the NTU-RGBD dataset, we use as baseline for comparison the features extracted with BoW method. As it can be seen in Table 4, our method gives a significant boost to the accuracy, especially in the multimodal case that is shown in column under av. Moreover, in Table 5, the case where one modality is missing, and thus, the classification is done for the three remaining modalities. Only the *Average training* approach is presented.

However, the most significant contributions of our method is that it is universal, it can be applied to any kind of data, and most importantly, it can deal with the cases of missing modalities. For example, columns under *Per modality CVA accuracy* (Tables 2, 4, 6) illustrate the method's performance if only one modality is available during the method testing,

and Table 5 shows the performance for one missing modality of NTU-RGBD dataset. Furthermore, objects with missing modalities can also be employed during training by following the *Separate training* approach.

5. Method Convergence and Computational cost

In this section we briefly present the computational cost of our method's time consuming parts, namely the Updating and the Optimization procedures as presented in Algorithms 1 and 2. We present these procedures since these are the most time consuming ones because they iterative procedures.

Updating: $2 \cdot 2 \cdot r \cdot [2O(n \cdot d \cdot s) + O(n \cdot s) + 3O(ld \cdot s)] \cdot \sum_c O((n_c - 1) \cdot n_c/2)$. Where $\sum_c O((n_c - 1) \cdot n_c/2)$ is the number of the different pairs of samples that are similar in terms of label (class). As n_c we consider the number of samples that belong to class c, n is the total samples and d and s the method's parameters (see Section 3.3). Parameter r is the maximum number of iterations of this part, the optimal value of which is found with heuristic method as described in Section 3.3, and the effect of it is presented in Figure 3a.

Optimization: $[5O(n'^2 \cdot l) + 2O(n'^3) + 7O(n'^2) + 4O(n' \cdot l^2) + O(l^3) + O(l^2)] \cdot rs$. Similarly, n' is the number of samples in the reshaped dataset, l the number of largest eigenvalues kept, and rs is the maximum number of iterations that can be performed before convergence, as explained in Sections 3.4 and 3.7. In this final step, the convergence of the method is measured using the Euclidean distance between the transformed data and the calculated embedding $\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2$. The method is iterating until the distance becomes smaller than a small predefined value, which has been heuristically estimated to be equal to 10^{-3}. The entire method is constrained at 200 repetitions, $r = 200$ (as referred in sub-section 4.1), a number that has been identified experimentally.

Indicatively, for the optimal parameter selections shown in bold in Table 1 and for the realization presented in subsections 4.1, 4.2 and 4.3, the mentioned parts were timed and the results are shown below.

1. In NUS-wide dataset, *Upd.* $\rightarrow$ 183, 02s and *Opt.* $\rightarrow$ 133, 45s (103 iterations until

convergence).

2. In NTU-RGBD dataset, *Upd.* $\rightarrow$ 32.66s and *Opt.* $\rightarrow$ 323.63s (200 iterations without convergence).
3. In AV-Letters, *Upd.* $\rightarrow$ 9.04s and *Opt.* $\rightarrow$ 95.44s (200 iterations without convergence).

The realizations took place in a machine with i5-6600 CPU @ 3.30Ghz and 16GB of installed RAM.

6. Conclusions

We have proposed a general-purpose, graph-based, multimodal fusion framework that can be used for multimodal data classification. This method is a combination of multimodal metric learning with a graph-based multimodal fusion method. The Bag of Words framework and neural networks have been used for feature extraction in the datasets in order to present results as dataset-independent as possible. Our method reaches classification accuracy of the state-of-the-art methods for classification of the single representation of a multimodal instance. It is very substantial that this method is able to use multimodal data with missing modalities in both the training and the testing procedures. Experiments in two well-known datasets proved that the proposed method outperforms other state-of-art methods in multimodal classification.

On the other hand, due to the fact that it is an iterative method, using SVD the proposed method cannot cope with large datasets of lots of modalities. However, since the results of the proposed method on small and medium size datasets are competitive, there should be more work on extending it in order to deal with large datasets by employing big data methods and tools to overcome specific bottlenecks such as the enormous dimensions of the Laplacian matrix of a graph of thousands of samples.

Appendix A. Proof of equation 16

Let

$$\mathcal{F} = \mathbf{Y}^T\mathbf{L_w}\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1} \tag{A1}$$

where $\mathbf{X} \in M^{(n\times d_1)}$ and $\mathbf{A} \in M^{(n\times d_2)}$

The three terms of (A1) are differentiated separately. The first term does not contain $\mathbf{A}$, thus it's derivative equals 0. Then the second term can be written as:

$$\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{d_2}(y_{ij} - \sum_{k=1}^{d_1} x_{ki}a_{kj})^2 \tag{A2}$$

hence its derivative equals:

$$\frac{\partial(\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2)}{\partial\mathbf{A}} = \sum_{i=1}^{n}\sum_{j=1}^{d_2}\frac{\partial((y_{ij} - \sum_{k=1}^{d_1} x_{ki}a_{kj})^2)}{\partial\mathbf{A}} \tag{A3}$$

Let $u = y_{ij} - \sum_{k=1}^{d_1} x_{ki}a_{kj}$ and $g(u) = u^2$.

Then

$$\frac{\partial((y_{ij} - \sum_{k=1}^{d_1} x_{ki}a_{kj})^2)}{\partial\mathbf{A}} = \frac{\partial g(u)}{\partial\mathbf{A}} = \frac{\partial g(u)}{\partial u}\cdot\frac{\partial u}{\partial\mathbf{A}} = 2u\cdot\frac{\partial u}{\partial\mathbf{A}} \tag{A4}$$

where

$$\frac{\partial u}{\partial\mathbf{A}} = \begin{bmatrix} \frac{\partial u}{\partial a_{11}} & \cdots & \frac{\partial u}{\partial a_{1j}} & \cdots & \frac{\partial u}{\partial a_{1d_2}} \\ & & \vdots & & \\ \frac{\partial u}{\partial a_{d_1 1}} & \cdots & \frac{\partial u}{\partial a_{d_1 j}} & \cdots & \frac{\partial u}{\partial a_{d_1 d_2}} \end{bmatrix} \tag{A5}$$

$$\frac{\partial u}{\partial a_{pq}} = -\sum_{k=1}^{d_1} x_{ki}\frac{\partial a_{kj}}{\partial a_{pq}} = -x_{pi} = -\begin{bmatrix} 0 & x_{1i} & 0 \\ 0 & \vdots & 0 \\ 0 & x_{d_1 i} & 0 \end{bmatrix} = -\mathbf{M}_j^{(i)}$$

So,

$$\frac{\partial(y_{ij} - \sum_{k=1}^{d_1} x_{ki}\cdot a_{kj})^2}{\partial\mathbf{A}} = -2(y_{ij} - \sum_{k=1}^{d_1} x_{ki}\cdot a_{kj})\cdot\mathbf{M}_j^{(i)} \tag{A6}$$

From (A3) and by setting $\mathbf{B} = \mathbf{Y} - \mathbf{X}^T\mathbf{A}$:

$$\begin{aligned}
\frac{\partial(\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2)}{\partial \mathbf{A}} &= -2\sum_{i=1}^{n}\sum_{j=1}^{d_2}(y_{ij} - \sum_{k=1}^{d_1} x_{ki} \cdot a_{kj}) \cdot \mathbf{M}_j^{(i)} \\
&= -2\sum_{i=1}^{n}\left[(y_{i1} - \sum_{k=1}^{d_1} x_{ki}a_{k1}) \cdot \mathbf{M}_1^{(i)} + (y_{id_2} - \sum_{k=1}^{d_1} x_{ki}a_{kd_2}) \cdot \mathbf{M}_{d_2}^{(i)}\right] \\
&= -2\sum_{i=1}^{n}\left[b_{i1}\mathbf{M}_1^{(i)} + \cdots + b_{id_2}\mathbf{M}_{d_2}^{(i)}\right] \\
&= -2\sum_{i=1}^{n}\begin{bmatrix} b_{i1}x_{1i} & \dots & b_{id_2}x_{1i} \\ & \vdots & \\ b_{i1}x_{d_1i} & \dots & b_{id_2}x_{d_1i} \end{bmatrix} \\
&= -2\mathbf{X} \cdot \mathbf{B} = -2\mathbf{X} \cdot (\mathbf{Y} - \mathbf{X}^T\mathbf{A}) = -2\mathbf{X}\mathbf{Y} + 2\mathbf{X}\mathbf{X}^T\mathbf{A}
\end{aligned} \tag{A7}$$

For the third term:

$$\frac{\partial\|\mathbf{A}\|_{2,1}}{\partial \mathbf{A}} = \sum_{i=1}^{d_1}\frac{\partial\sqrt{\sum_{j=1}^{d_2} a_{ij}^2}}{\partial \mathbf{A}} \tag{A8}$$

let $u(\mathbf{A}) = \sum_{j=1}^{d_2} a_{ij}$, $g(u) = \sqrt{u}$ and $a_i = \sum_{j=1}^{d_2} a_{ij}$

Therefore,

$$
\begin{aligned}
\frac{\partial \|\mathbf{A}\|_{2,1}}{\partial \mathbf{A}} &= \sum_{j=1}^{d_1} \frac{\partial g(u)}{\partial \mathbf{A}} = \sum_{j=1}^{d_2} \frac{\partial g}{\partial u} \cdot \frac{\partial u}{\partial \mathbf{A}} \\
&= \sum \frac{1}{2\sqrt{u}} \begin{bmatrix} \frac{\partial u}{\partial a_{11}} & \cdots & \frac{\partial u}{\partial a_{1d_2}} \\ & \vdots & \\ \frac{\partial u}{\partial a_{d_1 1}} & \cdots & \frac{\partial u}{\partial a_{d_1 d_2}} \end{bmatrix} = \sum_{i=1}^{d_1} \frac{1}{2\sum_{j=1}^{d_2} a_{ij}} \sum_{j=1}^{d_2} \begin{bmatrix} \frac{\partial a_{ij}^2}{\partial a_{11}} & \cdots & \cdots \\ \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \\
&= \sum_{i=1}^{d_1} \frac{1}{2\sqrt{\sum_{j=1}^{d_2} a_{ij}}} \sum_{j=1}^{d_2} \begin{bmatrix} 0 & \ldots & 0 & \ldots & 0 \\ 0 & \ldots & 2a_{ij} & \ldots & 0 \\ 0 & \ldots & 0 & \ldots & 0 \end{bmatrix} \\
&= \sum_{i=1}^{d_1} \|a_i\|^{-1} \begin{bmatrix} 0 & \ldots & 0 & \ldots & 0 \\ a_{i1} & \ldots & a_{ij} & \ldots & a_{id_2} \\ 0 & \ldots & 0 & \ldots & 0 \end{bmatrix} = \begin{bmatrix} \frac{a_{11}}{\|a_1\|} & \cdots & \frac{a_{1d_2}}{\|a_1\|} \\ \vdots & \ddots & \cdots \\ \frac{a_{d_1 1}}{\|a_{d_1}\|} & \cdots & \frac{a_{d_1 d_2}}{\|a_{d_1}\|} \end{bmatrix} = \boldsymbol{\Delta}\mathbf{A}
\end{aligned}
\tag{A9}
$$

$$
\text{where } \boldsymbol{\Delta} = \begin{bmatrix} \|a_1\|^{-1} & 0 & \ldots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \cdots & \|a_i\|^{-1} & 0 \\ 0 & \ldots & \ldots & \|a_{d_1}\|^{-1} \end{bmatrix} \tag{A10}
$$

By substituting to (A1), $\partial \mathcal{F}/\partial \mathbf{A}$ equals 0

$$
\begin{aligned}
\frac{\partial(\mathbf{Y}^T\mathbf{L}_\mathbf{w}\mathbf{Y} + \varpi\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2 + \sigma\|\mathbf{A}\|_{2,1}))}{\partial \mathbf{A}} &= 0 \Rightarrow \\
\frac{\varpi \cdot \partial(\|\mathbf{Y} - \mathbf{X}^T\mathbf{A}\|_2^2)}{\partial \mathbf{A}} + \frac{\sigma \cdot \partial(\|\mathbf{A}\|_{2,1}^2)}{\partial \mathbf{A}} &= 0 \Rightarrow \\
\varpi \cdot (-2\mathbf{X}\mathbf{Y} + 2\mathbf{X}\mathbf{X}^T\mathbf{A}) + \sigma \cdot \boldsymbol{\Delta}\mathbf{A} &= 0 \Rightarrow \\
-2\varpi\mathbf{X}\mathbf{Y} + (2\varpi\mathbf{X}\mathbf{X}^T + \sigma\boldsymbol{\Delta})\mathbf{A} &= 0 \Rightarrow \\
-\mathbf{X}\mathbf{Y} + (\mathbf{X}\mathbf{X}^T + \frac{\sigma}{2\varpi}\boldsymbol{\Delta})\mathbf{A} &= 0 \Rightarrow \\
\mathbf{A} = (\mathbf{X}\mathbf{X}^T + \frac{\sigma}{2\varpi}\boldsymbol{\Delta})^{-1} \cdot \mathbf{X}\mathbf{Y}&
\end{aligned}
\tag{A11}
$$

Acknowledgment

This work has been supported by the EU Horizon 2020 Framework Programme under grant agreement no. 690090 (ICT4Life project).

References

[1] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: An overview of methods, challenges, and prospects, Proceedings of the IEEE 103 (2015) 1449–1477.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696.

[3] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (2013) 28 – 44.

[4] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, arXiv preprint arXiv:1705.09406 (2017).

[5] A. Shahroudy, G. Wang, T.-T. Ng, Multi-modal feature fusion for action recognition in rgb-d sequences, in: Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on, IEEE, pp. 1–4.

[6] H. LI, J. Sun, X. Zongben, L. Chen, Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network, IEEE Transactions on Multimedia (2017).

[7] Z. Liu, L. Zhang, Q. Liu, Y. Yin, L. Cheng, R. Zimmermann, Fusion of magnetic and visual sensors for indoor localization: Infrastructure-free and more effective, IEEE Transactions on Multimedia 19 (2017) 874–888.

[8] M. Guillaumin, J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, pp. 902–909.

[9] J. Geng, Z. Miao, X.-P. Zhang, Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection, IEEE Transactions on Multimedia 17 (2015) 498–511.

[10] F. Sun, D. Harwath, J. Glass, Look, listen, and decode: Multimodal speech recognition with images, in: Spoken Language Technology Workshop (SLT), 2016 IEEE, IEEE, pp. 573–578.

[11] S. Receveur, D. Scheler, T. Fingscheidt, A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition, in: Situated Dialog in Speech-Based Human-Computer Interaction, Springer, 2016, pp. 179–192.

[12] D. Rafailidis, S. Manolopoulou, P. Daras, A unified framework for multimodal retrieval, Pattern Recognition 46 (2013) 3358–3370.

[13] S. Thermos, G. T. Papadopoulos, P. Daras, G. Potamianos, Deep affordance-grounded sensorimotor object recognition, margin 17 (2017) 35.

[14] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using rgb-d cameras, IEEE Transactions on Circuits and Systems for Video Technology 26 (2016) 788–799.

[15] F. Destelle, A. Ahmadi, N. E. O'Connor, K. Moran, A. Chatzitofis, D. Zarpalas, P. Daras, Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors, in: Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, IEEE, pp. 371–375.

[16] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[17] L. Xu, X. Wu, K. Chen, L. Yao, Multi-modality sparse representation-based classification for alzheimer's disease and mild cognitive impairment, Computer methods and programs in biomedicine 122 (2015) 182–190.

[18] E. A. Bernal, X. Yang, Q. Li, J. Kumar, S. Madhvanath, P. Ramesh, R. Bala, Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors, IEEE Transactions on Multimedia (2017).

[19] C. Tan, F. Sun, W. Zhang, J. Chen, C. Liu, Multimodal classification with deep convolutional-recurrent neural networks for electroencephalography, in: International Conference on Neural Information Processing, Springer, pp. 767–776.

[20] K. Kalimeri, C. Saitis, Exploring multimodal biosignal features for stress detection during indoor mobility, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, pp. 53–60.

[21] R. Wagh, S. Darokar, S. Khobragade, Multimodal biometrics features with fusion level encryption, International Journal of Engineering Science 5246 (2017).

[22] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, A. D. N. Initiative, et al., Multi-modal classification of alzheimer's disease using nonlinear graph fusion, Pattern Recognition 63 (2017) 171–181.

[23] A. Patwardhan, G. Knapp, Aggressive actions and anger detection from multiple modalities using kinect, arXiv preprint arXiv:1607.01076 (2016).

[24] F. Cricri, M. J. Roininen, J. Leppanen, S. Mate, I. D. Curcio, S. Uhlmann, M. Gabbouj, Sport type classification of mobile videos, IEEE Transactions on Multimedia 16 (2014) 917–932.

[25] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust rgb-d object recognition, in: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE, pp. 681–687.

[26] S. S. Mukherjee, N. M. Robertson, Deep head pose: Gaze-direction estimation in multimodal video, IEEE Transactions on Multimedia 17 (2015) 2094–2107.

[27] F. Gürpinar, H. Kaya, A. A. Salah, Multimodal fusion of audio, scene, and face features for first impression estimation, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, pp. 43–48.

[28] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, T.-S. Chua, Multi-label visual classification with label exclusive context, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, pp. 834–841.

[29] P. Xie, E. P. Xing, Multi-modal distance metric learning, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press, pp. 1806–1812.

[30] M. Zeppelzauer, D. Schopfhauser, Multimodal classification of events in social media, Image and Vision Computing 53 (2016) 45–56.

[31] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, H. Sawhney, Multimodal fusion using dynamic hybrid models, in: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, IEEE, pp. 556–563.

[32] K. Sohn, W. Shang, H. Lee, Improved multimodal deep learning with variation of information, in: Advances in Neural Information Processing Systems, pp. 2141–2149.

[33] J. Li, Y. Wu, J. Zhao, K. Lu, Multi-manifold sparse graph embedding for multi-modal image classification, Neurocomputing 173 (2016) 501–510.

[34] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE transactions on pattern analysis and machine intelligence 37 (2015) 2085–2098.

[35] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE Transactions on Multimedia 17 (2015) 1989–1999.

[36] J. Li, Zechao Tang, Weakly supervised deep matrix factorization for social image understanding, IEEE Transactions on Image Processing 26 (2017) 276–288.

[37] P. C. Hansen, The truncatedsvd as a method for regularization, BIT Numerical Mathematics 27 (1987) 534–553.

[38] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM international conference on image and video retrieval, ACM, p. 48.

[39] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019.

[40] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 198–213.

[41] D. Hu, X. Li, et al., Temporal multimodal learning in audiovisual speech recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3574–3582.