A KNOWLEDGE REPRESENTATION FRAMEWORK FOR INTEGRATION, CLASSIFICATION OF MULTI-SCALE IMAGING AND NON-IMAGING DATA: PRELIMINARY RESULTS IN PREDICTING PROSTATE CANCER RECURRENCE BY FUSING MASS SPECTROMETRY AND HISTOLOGY

George Lee, Scott Doyle, James Monaco, Anant Madabhushi *

Rutgers University
Dept of Biomedical Engineering
Piscataway, New Jersey, 08854, USA

Michael D. Feldman, Stephen R. Master, John E. Tomaszewski

University of Pennsylvania
Department of Pathology
Philadelphia, Pennsylvania, 19104, USA

ABSTRACT

The demand for personalized health care requires a wide range of diagnostic tools for determining patient prognosis and theragnosis (response to treatment). These tools present us with data that is both multi-modal (imaging and non-imaging) and multi-scale (proteomics, histology). By utilizing the information in these sources concurrently, we expect significant improvement in predicting patient prognosis and theragnosis. However, a prerequisite to realizing this improvement is the ability to effectively and quantitatively combine information from disparate sources. In this paper, we present a general fusion framework (GFF) aimed towards a combined knowledge representation predicting disease recurrence. To the best of our knowledge, GFF represents the first formal attempt to fuse biomedical image and non-image information directly at the data level as opposed to the decision level, thus preserving more subtle contributions in the original data. GFF represents the different data streams in separate embedding spaces via the application of dimensionality reduction (DR). Data fusion is then implemented by combining the individual reduced embedding spaces. A proof of concept example is considered for evaluating the GFF, whereby protein expression measurements from mass spectrometry are combined with histological image signatures to predict prostate cancer (CaP) recurrence in 6 CaP patients, following therapy. Preliminary results suggest that GFF offers an intelligent way to fuse image and non-image data structures for making prognostic and theragnostic predictions.

Index Terms— Knowledge representation, data fusion, mass spectrometry, histopathology, prostate cancer, prognosis

1. INTRODUCTION

There exists an increasing demand for personalized medicine, a system which utilizes a patient's unique physiological and genetic profile to create tailor-made treatment plans. Utilizing a comprehensive patient profile reduces the likelihood of treatment failure because patients can be matched up with similar studies showing successful treatment. These characteristics will be captured via such disparate data sources as mass spectrometry, radiologic imaging, histology, and gene expression. Thus, the success of personalized medicine will depend greatly upon our ability to integrate multi-modal, multi-scale and multi-protocol data since 1) each individual source may contain information unavailable in the others, and 2) important dependencies between the sources can only be identified when they are considered concomitantly. Unfortunately, the fusion of dissimilar data is a non-trivial task. For example, consider the complications with fusing magnetic resonance (MR) imaging (structural information) in the form of scalar intensity information with MR spectral data (metabolic information) in vectorial form; each modality encodes different types of information, at different scales. Nonetheless, both modalities reflect information regarding the same disease, and consequently, examining both concurrently is crucial.

Previous attempts at image and non-image fusion have been geared toward using non-imaging information to aid in object detection, segmentation, and tracking in images, while data fusion for the purpose of classification has yet to be fully explored. Currently, information fusing algorithms are categorized as being either combination of data (COD) or combination of interpretations (COI) [1] methodologies. COD algorithms propose fusion at the feature level followed by classification. Mandic [2] proposed a COD methodology that combined heterogeneous wind speed and directional data using vector spaces of complex numbers. Lanckriet [3] combined amino acid sequences, protein complex data, gene expression data, and protein interactions directly in a kernel space to predict the functions of yeast proteins. COD methods (Figure 1(a)) aggregate features from each source into a single feature vector before classification. This has the advantage of retaining any inter-source dependencies between features. However, COD methods suffer from the curse-of-dimensionality (too many features), and consequently, require a vast amount of training data to produce a classifier that is extensible beyond the training set. Additionally, aggregating data from very different sources without accounting for differences in the number of features and their relative scalings can negatively impact certain classifiers.

Alternatively, COI methods classify the data from each source independently and then aggregate the results. Rohlfing [1] applied COI methods to combine different information sources for the following applications: atlas-based segmentation, image segmentation, and deformation-base morphometry. Jesneck [4] combined human- and computer-extracted mammographic feature sets by first classifying them independently and then fusing the binary decisions. Since

*This work made possible via grants from Coulter Foundation (WHCF 4-29368), New Jersey Commission on Cancer Research, National Cancer Institute (R21CA127186-01, R03CA128081-01), the Life Science Commercialization Award, and the US Department of Defense (W81XWH-08-1-0145).

each individual classifier produces a one-dimensional output in COI algorithms (shown in Figure 1(b)), the curse-of-dimensionality can be greatly mitigated. Furthermore, the classification results are implicitly normalized, facilitating the following classification step. However, all inter-source dependencies among features are lost.

Based on the above, it is apparent that both COI and COD approaches have inherent drawbacks that are difficult to overcome. COD methods suffer the curse-of-dimensionality while COI methods are unable to leverage inter-source dependencies. In this paper we suggest that the COI/COD dichotomy is a limited paradigm. It is more appropriate to consider a general fusion framework (GFF) in which COD and COI exist as two extremes of continuous spectrum. That is, COI methods reduce the feature sets from each source to a single dimension (by classification) and then aggregate the classification results; COD methods aggregate the features and then classify them. We propose a knowledge representation scheme that transforms, instead of classifies, the disparate feature sets. This hybridization produces new features, which if the transformations are chosen correctly, are of lower dimensionality and yet still retain all essential class discriminatory information. Consequently, GFF offers a usable combined representation of very different types of data, mitigating the typical drawbacks inherent in COD and COI algorithms.

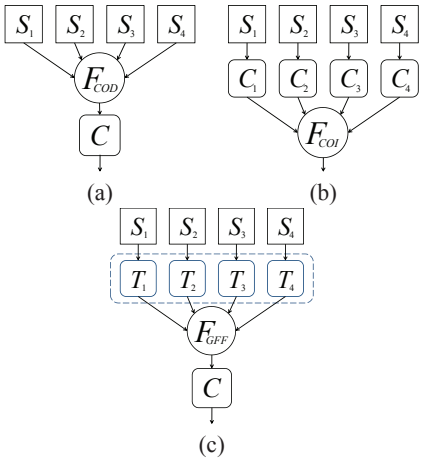


Fig. 1. (a) COD model: data from disparate sources (S_1-S_4) is first aggregated and then classified. (b) COI model: data is classified and then aggregated. (c) Generalized Fusion Framework: data from individual sources is transformed into a common knowledge framework then aggregated and transformed into a final interpretation.

To further elucidate our framework we will consider the specific example of fusing image-based features derived from hematoxylin and eosin (H&E) stained prostate tissue samples with peptide measurements derived from Electrospray Ionization Mass Spectrometry (ESI-MS). We demonstrate how GFF can be applied to uniformly represent information from these disparate sources and present preliminary results (intended to show proof of concept) in constructing an integrated meta-classifier for predicting CaP recurrence/non-recurrence from 6 patients previously treated for CaP.

2. GENERALIZED FUSION FRAMEWORK (GFF)

Our raw data is defined in terms of source $S_i(x_1, x_2, ..., x_k)$, where $x_1, x_2, ..., x_k$ represent the k observations in the study and i represents one of N data sources $i \in \{1, 2, ..., N\}$, omitting the $x_1, x_2, ..., x_k$ in our notation for convenience. GFF (Figure 1(c)) begins by applying separate transformations to each of the N data sources S_i, mapping them into a common knowledge representation T_i, $i \in \{1, 2, \ldots, N\}$ (shown by the dashed box). The features resulting from this transformation T_i are then aggregated into a fused space F prior to a classification C. Notice that if each transformation T_i were chosen to be classifier outputs, i.e. $T_i = C_i$, this would devolve to the COI model. Conversely, if each T_i simply passed the data without modification, the COD model would result.

2.1. Data Transformation: Meta-Space Projection (T_i)

The most important attribute of the GFF is its flexibility with respect to the choice of transformations. This flexibility allows us to tailor each transformation to best accommodate its individual source. The goal of the selected transformations is to map the data sources into a common meta-space that preserves inter-source dependencies, mitigates the curse-of-dimensionality, and is appropriate for visualization and analysis.

Consider the use of dimensionality reduction (DR) methods as transformations. Having the ability to distill datasets to a few informative features (i.e. their intrinsic dimensionality)[1], DR techniques can accomplish the goal of minimizing the curse-of-dimensionality and retaining inter-source dependencies. Since the GFF does not espouse any single DR technique, but instead contains them all, different DR algorithms can be applied to different sources. For example, we have shown that many types of medical data lie on nonlinear manifolds and are more amenable to manifold learning techniques such as LLE or ISOMAP [5]. For other sources, or for sparsely packed datasets, linear DR methods such as Principal Component Analysis (PCA) may be more appropriate.

2.2. Data Integration: Fusion of Multi-modal Data (F)

Following data transformation, the modalities S_i are now in a transformed space T_i that is more amenable for integration. To minimize bias from utilizing data of multiple scales, the DR transformed meta-spaces must first be normalized. We use the normalization

$$\mathcal{T}_i(x_a) = \frac{T_i(x_a) - \min_b[T_i(x_b)]}{\max_b[T_i(x_b)] - \min_b[T_i(x_b)]}, a, b \in \{1, 2, ..., k\} \quad (1)$$

for each T_i, $i \in \{1, 2, ..., N\}$. We then concatenate the modalities represented by the normalized transformations $\mathcal{T}_i$ to form fusion space $F = [\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N]$, which can be reduced again by DR.

3. PREDICTING CAP RECURRENCE BY FUSING MASS SPECTROMETRY AND HISTOPATHOLOGY

Prostate cancer is the most commonly diagnosed cancer among men in the United States, with an incidence of about 200,000 a year *(Source: American Cancer Society)*. Adequate cancer stratification can provide improvements in patient prognosis and theragnosis. Thus, many studies have been done that use either imaging methods or gene and protein expression to improve cancer stratification [6].

Our dataset consists of six patients from the Hospital of the University of Pennsylvania. From each of the six patients, prostate whole-mount histological slices (WMHSs) with corresponding mass

[1]Note that there are several algorithms that can be used to estimate intrinsic dimensionality, prior to meta-space projection

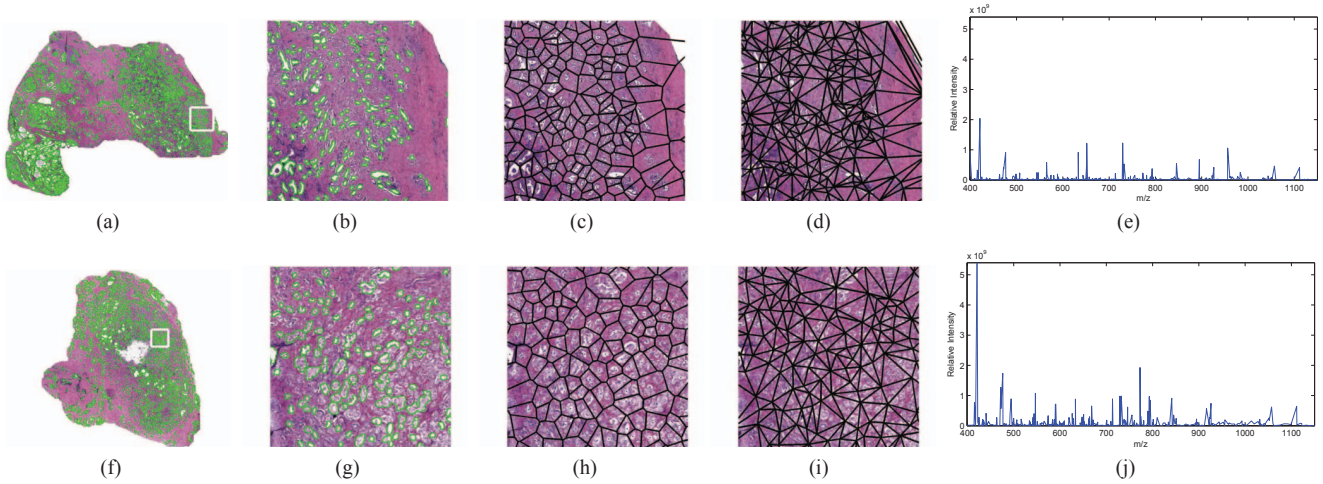


Fig. 2. Multi-modal patient data (top row: relapsed case, bottom row: non-relapsed case). (a), (f) Original whole-mount H&E prostate histology section showing region of interest. (b), (g) Magnified ROI showing gland segmentation boundaries, as well as the (c), (h) Voronoi Diagram and (d), (i) Delaunay Triangulation depicting gland architecture. (e), (j) Plot of the proteomic mass spectra profile.

spectrometry signals were obtained at the Department of Surgical Pathology. All patients in this study have been classified either as a relapsed (R) case, where cancer recurrence was observed following treatment, or a non-relapsed (NR) case, where cancer was not observed. For all 6 patients, CaP treatment involved radical prostatectomy. The motivation of this study was to observe whether the quantitative integration of image-based signatures acquired from the histological whole-mount prostate sections with corresponding peptide measurements obtained from mass spectrometry could be used to discriminate the 3 CaP progressors from the 3 non-progressors.

3.1. Image Feature Extraction from Prostate Histopathology

In prostate whole-mount histology (Figure 2 (a), (f)), the objects of interest are the glands (shown in Fig. 2 (b), (g)), whose shape and arrangement are highly correlated with cancer progression [7]. We briefly describe this process below. Prior to extracting image features, we employ an automatic region-growing gland segmentation algorithm presented by Monaco et al. [8]. The boundaries of the interior gland lumen (see Fig. 2(b)) and the centroids of each gland, allow for extraction of 1) morphological and 2) architectural features from histology as described in [7] and also briefly below.

Morphological Features (MF): We extract 25 morphological features (20 boundary features and 5 gland area features) from each of the glands within an image and calculate the average, median, standard deviation, and min-to-max ratio of the values for each feature to obtain 100 morphological features. These features provide information such as boundary smoothness, gland area, moment invariants, Fourier descriptors, and fractal dimensions.

Architectural Features (AR): We quantify the arrangement of the glands using a graph-based approach developed in [7]. Using the centroids of the glands as vertices, we construct three graphs: the Voronoi Diagram, the Delaunay Triangulation, and the Minimum Spanning Tree. By measuring statistics related to these graphs, such as average branch length, polygon or triangular area and perimeter, we extract 26 graph-based features. In addition, we calculate a set of architectural features from the spatial arrangement of the centroids, such as density and compactness, for a total set of 51 architectural features per image.

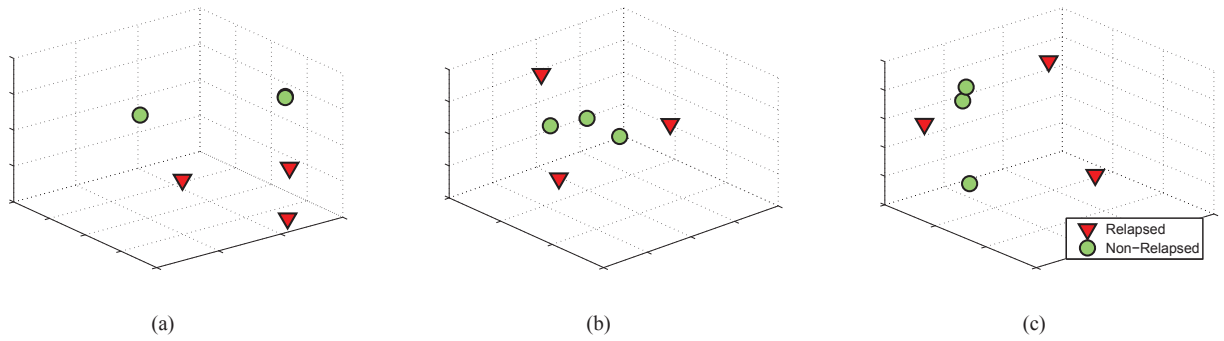
3.2. Peptide Measurements via Mass Spectrometry (MS)

In addition to the gold standard of histopathology, recent attempts have been made to identify a set of biological markers that can predict whether a patient is susceptible to cancer progression and recurrence [6]. Active genes encode proteins that are present in a tissue sample, and these proteins can be measured and serve as quantitative markers of cancer activity. For this study, we used Electrospray Ionization Mass Spectrometry (ESI-MS) to measure the relative abundance of peptides (expressed as mass/charge or m/z ratios) in cancerous regions of the tissue. Samples of formalin-fixed, paraffin-embedded (FFPE) prostate tissue measuring approximately 4 millimeters in diameter are digested into a protein lysate, resulting in the isolation of 8 micro grams of peptide per sample. This material is separated through high-performance liquid chromatography (C-18 reverse phase) and injected into a high-resolution, accurate-mass hybrid ion trap mass spectrometer (Thermo LTQ Orbitrap), where peptide features are identified from ESI-MS scans using the Hardklor/Kronik package [9].

In our study, in-house software was used to identify levels corresponding to 11,752 m/z features for each of the six patients, producing a high-dimensional feature vector characterizing each patient's protein expression profile at the time of treatment. The summed average of peaks were used to characterize the patient profile, using only peaks that could be used for all patients in our study, resulting in a final 570 m/z values used for data fusion.

3.3. Fusion of Image and Non-Image Data

We provide 3 modalities S_1-S_3 for our patient data: Architectural features (AR) of dimensionality $\mathbb{R}^{51}$ and Morphological features (MF) of dimensionality $\mathbb{R}^{100}$ extracted from WMHSs and m/z values from mass spectrometry (MS) of dimensionality $\mathbb{R}^{570}$. These are then independently transformed into a common low-dimensional meta-space projection T_i, $i \in \{1, 2, 3\}$ of dimensionality $\mathbb{R}^3$. In



(a) (b) (c)

Fig. 3. Low-dimensional representations of patient data following application of GFF using PCA. Red triangles indicate Relapsed cases, while green circles represent Non-Relapsed cases, described using (a) Mass Spectrometry features, (b) Histological (Morphological + Architectural) features, and (c) Multi-modal fusion of Morphological, Architectural, and Mass Spectrometry features.

our GFF implementation, Principal Component Analysis (PCA) is used to perform the transformation from S_i to T_i. PCA is a linear DR method that reduces the data to dominant eigenvectors that can be used to represent high dimensional data. Prior to data fusion, we also normalize the data for each T_i, $i \in \{1, 2, 3\}$ via Equation 1. This normalized meta-space data $\mathcal{T}_i$ is then concatenated to form fused data F of dimensionality $\mathbb{R}^9$ prior to a second reduction by PCA to a final dimensionality $\mathbb{R}^3$, representing a low-dimensional representation of the fused multi-modal CaP data F.

4. PRELIMINARY RESULTS AND DISCUSSION

To determine the interplay between data fusion and data features on prostate cancer detection, we test out our GFF on the following combination of features: 1) MS, 2) MF + AR, 3) MS + MF + AR. Figures 3(a), (b), and (c) show the transformations of the mass spectrometry (MS), histological (MF+AR), and a combination of both (MS+MF+AR) into a common space that is ideal for analysis and visualization. The combined meta-space in Figure 3(c) illustrates the representation of the patients as determined by both image (i.e. histology) and non-image (mass spectrometry) feature types. This is a unique way of integrating two disparate types of features and can easily be extended to additional information sources independent of the modality. These preliminary results reflect the applicability of GFF in fusing imaging and non-imaging information for discriminating between patients with different prognostic and theragnostic disease profiles.

5. CONCLUDING REMARKS

The utility of both traditional histology and mass spectrometry has the chance to produce truly landscape-altering research in the near future, and the successful fusion of these information rich modalities will be vital for future research in successfully predicting disease prognosis and theragnosis studies. In this paper we have introduced a novel general fusion framework (GFF) for knowledge representation of data from imaging and non-imaging sources that is both easily implemented and extensible. While previous work has focused on decision-level fusion, our method combines these disparate sources at the data level while avoiding direct data concatenation, thus avoiding the drawbacks associated with the COI and COD methods. In spite of a small sample size, we have been able to show via preliminary results of the efficacy of our knowledge representation framework for fusing disparate data types for classification. Future work utilizing the GFF may explore other DR methods beyond PCA (including non-linear schemes), the optimal dimensionality of each data source, additional implementations of meta-space fusion apart from concatenation, and expansion to the fusion of several disparate data sources beyond histopathology and peptide features.

6. REFERENCES

[1] T. Rohlfing et al., "Information fusion in biomedical image analysis: Combination of data vs. combination of interpretations," *IPMI*, vol. 3565, pp. 150–161, 2005.

[2] D.P. Mandic et al., "Sequential data fusion via vector spaces: Fusion of heterogeneous data in the complex domain," *J. VLSI Sig. Proc. Syst.*, vol. 48, no. 1-2, pp. 99–108, 2007.

[3] G.R.G. Lanckriet et al., "Kernel-based data fusion and its application to protein function prediction in yeast.," *Pac. Symp. Biocomp.*, pp. 300–311, 2004.

[4] J.L. Jesneck et al., "Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis.," *Med Phys*, vol. 33, no. 8, pp. 2945–2954, Aug 2006.

[5] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE/ACM Trans. on Comp. Biology and Bioinf.*, vol. 5, no. 3, pp. 368–384, 2008.

[6] M.E. Wright et al., "Mass spectrometry-based expression profiling of clinical prostate cancer.," *Mol Cell Proteomics*, vol. 4, no. 4, pp. 545–554, Apr 2005.

[7] S. Doyle et al., "Using manifold learning for content-based image retrieval of prostate histopathology," in *MICCAI*, 2007.

[8] J.P. Monaco et al., "Detection of prostate cancer from whole-mount histology images using markov random fields," in *MIAAB*, 2008.

[9] M.R. Hoopman et al., "High-speed data reduction, feature detection, and ms/ms spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry," *Anal Chem*, vol. 79, pp. 5620–5632, 2007.