Weighted Linear Fusion of Multimodal Data – A Reasonable Baseline?

Ognjen Arandjelović
School of Computer Science
University of St Andrews
St Andrews KY16 9SX
United Kingdom
ognjen.arandjelovic@gmail.com

ABSTRACT

The ever-increasing demand for reliable inference capable of handling unpredictable challenges of practical application in the real world, has made research on information fusion of major importance. There are few fields of application and research where this is more evident than in the sphere of multimedia which by its very nature inherently involves the use of multiple modalities, be it for learning, prediction, or human-computer interaction, say. In the development of the most common type, score-level fusion algorithms, it is virtually without an exception desirable to have as a reference starting point a simple and universally sound baseline benchmark which newly developed approaches can be compared to. One of the most pervasively used methods is that of weighted linear fusion. It has cemented itself as the default off-the-shelf baseline owing to its simplicity of implementation, interpretability, and surprisingly competitive performance across a wide range of application domains and information source types. In this paper I argue that despite this track record, weighted linear fusion is not a good baseline on the grounds that there is an equally simple and interpretable alternative – namely quadratic mean-based fusion – which is theoretically more principled and which is more successful in practice. I argue the former from first principles and demonstrate the latter using a series of experiments on a diverse set of fusion problems: computer vision-based object recognition, arrhythmia detection, and fatality prediction in motor vehicle accidents.

Keywords

Prediction; arrhythmia; object recognition; computer vision; car accident.

1. INTRODUCTION

Score-level fusion of information is pervasive in algorithms addressing a wide variety of problems. From predictions of the price of French vintage wine using the fusion of predictions based on rainfall and temperature data [13] or more robust estimates of a person's gaze through the fusion of estimates obtained by considering each of the eyes in isolation [10], to sophisticated biometric algorithms which fuse similarity measures based on visual and infrared face appearance [11, 6] or gait characteristics [14], the fundamental premise underlying information fusion is the same: the use of multiple information sources facilitates the making of better decisions, be they in the form of classification, regression, or any of a number of others.

Whatever the specific problem, the effectiveness of a specific information fusion methodology needs to be demonstrated. If possible this should be done by comparing its performance against the current state-of-the-art. However this is often prohibitively difficult in practice. For example, in some cases the state-of-the-art may not be readily available for evaluation – the original code may not have been released and re-implementation may be overly time-consuming, the algorithm may not have been described in sufficient detail, the technology may be proprietary and costly to purchase, etc. In other cases there may not be a clear state-of-the-art because the particular problem at hand has not been addressed before. In such circumstances it is useful to compare the novel methodology with a simple yet sensible baseline.

Probably the simplest choices for the baseline would be the performances achieved using individual information sources which are being fused. However this is an excessively low bar for comparison. Instead what is widely done by authors across the research spectrum is to use simple weighted fusion of individual scores as a reference.

2. UNINFORMED INFORMATION FUSION

Let us start by formalizing the problem considered in this paper. I adopt the broad paradigm of so-called data source identification and matching. In particular, I assume that the available data comprises a set of n sensed observations, $\mathbf{X} = \{\mathbf{x}_i\} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, which correspond to different modalities sensing the same data source. Furthermore, I assume that there are functions (i.e. algorithms) $\phi_1, \ldots, \phi_n$ which quantify how well a particular observation $\mathbf{x}_i$ matches a specific data source. Without loss of generality I assume $0 \leq \phi_i(\mathbf{x}_i) < \infty$, where 0 indicates the best possible match, and ∞ the worst i.e. $\phi_i(\mathbf{x}_i)$ can be thought of a quasi-distance – 'quasi-' to emphasise that it is not required that $\phi_i(\mathbf{x}_i)$ meets the strict conditions required of a metric. Note that the aforestated setting describes observations in the most general sense, e.g. these may be feature vectors (such as rasterized appearance images or SIFT descriptors in computer vision), sets of vectors, and so on, and they do not need to be of the same type (for example, some may be vectors, others sequences) or dimensionalities.

As noted in the previous section, we are looking for a simple way of fusing different matching decisions $\phi_i(\mathbf{x}_i)$ in a manner which provides a reasonable baseline in an uninformed setting, that is,

without exploiting any particular properties of different functions ϕ_i or observations $\mathbf{x}_i$. A common one pervasively used in the literature is weighted linear fusion which can be expressed as follows:

$$\Phi(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n) = \sum_{i=1}^{n} \left[w_i \times \phi_i(\mathbf{x}_i) \right], \tag{1}$$

where:

$$\forall i = 1,\ldots,n.\ w_i \geq 0, \qquad \text{and} \qquad \sum_{i=1}^{n} w_i = 1. \tag{2}$$

Here $\Phi(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n)$ is the fused matching score which is derived by simply combining different $\phi_i(\mathbf{x}_i)$, and (2) ensures that the condition $0 \leq \phi_i(\mathbf{x}_i) < \infty$ is maintained for $\Phi(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n)$ as well, i.e. that it is the case that $0 \leq \Phi(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n) < \infty$.

The fusion approach described by (1) can also be rewritten more simply as:

$$\Phi(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n) = \tag{3}$$

$$\hat{\Phi}(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,w_1,\ldots,w_n) = \tag{4}$$

$$\sum_{i=1}^{n} \left[\frac{1}{n} \times \hat{\phi}_i(\mathbf{x}_i, w_i, n) \right], \tag{5}$$

where:

$$\hat{\phi}_i(\mathbf{x}_i, w_i, n) = n \times w_i \times \phi_i(\mathbf{x}_i). \tag{6}$$

Here the original weighted linear fusion of the contributing terms $\phi_i(\mathbf{x}_i)$ in (1) has been replaced by unweighted linear fusion of terms $\hat{\phi}_i(\mathbf{x}_i, w_i, n)$ which adjust the magnitudes of the fused scores directly. This can be achieved simply by considering the statistics of the distributions of different $\phi_i(\mathbf{x}_i)$ and the corresponding prediction, and without any knowledge of what different features $\mathbf{x}_i$ represent or what form different ϕ_i have, i.e. while maintaining the premise of uninformed fusion.

It can be readily seen that the described fusion approach fits the conditions specified in the preceding section – namely, simplicity (both methodological and that of implementation) and broad applicability. Although simple, uninformed fusion demonstrates remarkably good performance across a wide span of different data types and domains, ranging from dementia screening [17] to multimodal biometric identification [14, 5, 12].

$$\Phi_o(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,\omega_1,\ldots,\omega_n) = \tag{7}$$

$$\sqrt{\sum_{i=1}^{n} \left[\omega_i \times \phi_i(\mathbf{x}_i)^2 \right]}, \tag{8}$$

where as before:

$$\forall i = 1,\ldots,n.\ \omega_i \geq 0, \qquad \text{and} \qquad \sum_{i=1}^{n} \omega_i = 1. \tag{9}$$

$$\Phi_o(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,\omega_1,\ldots,\omega_n) = \tag{10}$$

$$\hat{\Phi}_o(\mathbf{x}_1,\ldots,\mathbf{x}_n;\phi_1,\ldots,\phi_n,\omega_1,\ldots,\omega_n) = \tag{11}$$

$$\sqrt{\sum_{i=1}^{n} \left[\frac{1}{n} \times \widetilde{\phi}_i(\mathbf{x}_i, \omega_i, n)^2 \right]}, \tag{12}$$

where, analogously to (6), the adjusted fusion scores can be expressed as:

$$\widetilde{\phi}_i(\mathbf{x}_i, \omega_i, n) = \sqrt{n} \times \omega_i \times \phi_i(\mathbf{x}_i). \tag{13}$$

The formulation in (12) can be readily recognized as the quadratic mean, in engineering also commonly referred to as the root mean square (RMS), of the n terms $\widetilde{\phi}_i(\mathbf{x}_i, \omega_i, n)$.

Before proceeding with a formal analysis of the two fusion approaches, it is insightful to gain an intuitive understanding of the difference between them. Note that simple linear fusion described in (5) treats different scores as effectively interchangeable – a decrease in one score is exactly compensated by an increase in another score by the same amount. This is sensible when the scores correspond to the same measurement which is merely repeated. The described fusion can then be seen as a way of reducing measurement error, assuming that measurement is unbiased and that errors are identically and independently distributed [9]. However this is a rather trivial case of fusion and in practice one is more commonly interested in fusion of different types of data modalities. Indeed, in practice it is often the explicit aim to try to use information sources which vary independently, and attempt to exploit best their complementary natures [5]. In such instances different scores are best treated as describing a source in orthogonal directions, thereby giving rise to a feature vector $\left[\widetilde{\phi}_1(\mathbf{x}_1),\ldots,\widetilde{\phi}_n(\mathbf{x}_n) \right]^T \in \mathbb{R}^n$. The quadratic mean-based fusion of (12) can then be thought of as emerging from a normalized distance measure between such feature vectors in the corresponding ambient embedding space $\mathbb{R}^n$.

Let us now consider the effects of the two fusion approaches in more detail. Firstly, note that scoring measures $\hat{\phi}_i(\mathbf{x}_i, w_i, n)$ and $\widetilde{\phi}_i(\mathbf{x}_i, \omega_i, n)$ are inevitably imperfect – neither can be expected to produce universally the perfect matching score of 0 when the query correctly matches a source and ∞ when it does not. Even the weaker requirement of universally smaller pseudo-distances for correct matches is unrealistic. Indeed, it is precisely this practical challenge that motivates multimodal fusion. Consequently, they are appropriately modelled as resulting from draws from random variables:

$$\hat{\phi}_i(\mathbf{x}_i), \widetilde{\phi}_i(\mathbf{x}_i) \sim X_i. \tag{14}$$

Let us examine the effects of the two fusion approaches in detail. Specifically, for clarity consider the fusion of two sources, say i and j, while observing that the derived results are readily applicable to the fusion of an arbitrary number of sources through the use of an inductive argument. Following (5) and (12), applying linear and quadratic mean-based fusion results in scores described respectively by random variables $\hat{Y}_{ij}$ and $\widetilde{Y}_{ij}$, where:

$$\frac{\hat{\phi}_i(\mathbf{x}_i) + \hat{\phi}_j(\mathbf{x}_j)}{2} \sim \frac{X_i + X_j}{2} = \hat{Y}_{ij}, \tag{15}$$

and:

$$\sqrt{\frac{\widetilde{\phi}_i(\mathbf{x}_i)^2 + \widetilde{\phi}_j(\mathbf{x}_j)^2}{2}} \sim \sqrt{\frac{X_i^2 + X_j^2}{2}} = \widetilde{Y}_{ij}. \tag{16}$$

The former can be further expanded as:

$$E\left[\hat{Y}_{ij}^2 \right] = E\left[\left(\frac{X_i + X_j}{2} \right)^2 \right] \tag{17}$$

$$= \frac{1}{4} \left\{ E\left[X_i^2 \right] + E\left[X_j^2 \right] + E\left[X_i X_j \right] \right\}, \tag{18}$$

and the latter as:

$$E\left[\widetilde{Y}_{ij}^2\right] = E\left[\sqrt{\frac{X_i^2 + X_j^2}{2}}^2\right] \tag{19}$$

$$= \frac{1}{2}\left\{E\left[X_i^2\right] + E\left[X_j^2\right]\right\} \tag{20}$$

$$= \frac{1}{4}\left\{2E\left[X_i^2\right] + 2E\left[X_j^2\right]\right\}. \tag{21}$$

Writing:

$$\mu_i \equiv E\left[X_i\right], \tag{22}$$

and

$$\mu_j \equiv E\left[X_j\right], \tag{23}$$

as well as:

$$\sigma_i \equiv E\left[(X_i - \mu_i)^2\right], \tag{24}$$

and

$$\sigma_j \equiv E\left[(X_j - \mu_j)^2\right], \tag{25}$$

and using the standard definition of Pearson's correlation coefficient ρ:

$$\rho = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_i\,\sigma_j}, \tag{26}$$

we can write:

$$E\left[\hat{Y}_{ij}^2\right] - E\left[\widetilde{Y}_{ij}^2\right] = \tag{27}$$

$$\sigma_i^2 + \mu_i^2 + \sigma_j^2 + \mu_j^2 - 2\left(\rho\sigma_i\sigma_j - \mu_i\mu_j\right) = \tag{28}$$

$$\sigma_i^2 + \sigma_j^2 - 2\rho\sigma_i\sigma_j + (\mu_i + \mu_j)^2 \geq \tag{29}$$

$$\sigma_i^2 + \sigma_j^2 - 2\sigma_i\sigma_j + (\mu_i + \mu_j)^2 = \tag{30}$$

$$(\sigma_i + \sigma_j)^2 + (\mu_i + \mu_j)^2 \geq 0. \tag{31}$$

Therefore: $E\left[\hat{Y}_{ij}^2\right] \geq E\left[\widetilde{Y}_{ij}^2\right]$. In other words, following the fusion of the same scores the proposed quadratic mean-based fusion results in lower fused matching scores (quasi-distances described by the random variable $\widetilde{Y}_{ij}$) than those obtained by employing linear fusion (described by the random variable $\hat{Y}_{ij}$).

At first sight, the significance of the finding in (31) is not clear, given that it applies equally to the fusion of scores which result from matching and non-matching sources – quadratic mean-based fusion produces lower fusion scores in both cases. For the proposed fusion strategy to be advantageous it has to exhibit a differential effect and reduce matching scores more than non-matching ones. To see why there are indeed sound reasons to expect this to be the case, consider the intermediate result in (29). From this expression it can be readily seen that the reduction in the magnitude of the fused score effected by the proposed quadratic mean-based fusion in comparison with the linear baseline is dependent on Pearson's correlation ρ between random variables X_i and X_j which capture the stochastic properties of the original non-fused scores. Specifically, the greater the correlation between them the greater the corresponding reduction in the fused score becomes. The reason why this observation is key lies in the nature of the problem at hand: though imperfect, by their very design sensible matching functions which produce different $\hat{\phi}_i$ and $\widetilde{\phi}_i$ ($i = 1\dots n$) should be expected preferentially and systematically to produce lower scores for correctly matching sources and higher scores for incorrectly matching sources. Therefore, while non-matching comparisons may result in

some of the fused scores erroneously to be low, on average such errors should exhibit a lower degree of correlation than low scores across different modalities do for correct matches. In the next section I will demonstrate that this indeed is the case in practice.

3. EXPERIMENTS

Having laid out the theoretical argument against simple weighted combination as the default baseline for uninformed fusion, in favour of the quadratic mean, in this section the two are compared empirically on three real, challenging data sets of vastly different types. In particular, the popular computer vision problem of image-based object recognition is used as a case study in Section 3.2, arrhythmia prediction in Section 3.3, and finally fatality prediction in road vehicle accidents in Section 3.4.

3.1 Performance evaluation

I compared the performance of the two fusion approaches discussed in Section 2 by computing fused quasi-similarities between the ground truth data and the two information sources. Specifically, the linearly fused quasi-similarity of $\hat{\phi}_1^{(1)}(x_{i,1})$ and $\hat{\phi}_1^{(1)}(x_{i,2})$ was computed as:

$$\frac{1}{2}\left(\hat{\phi}_1^{(1)}(x_{i,1}) + \hat{\phi}_1^{(1)}(x_{i,2})\right), \tag{32}$$

and similarly of of $\hat{\phi}_1^{(2)}(x_{i,1})$ and $\hat{\phi}_1^{(2)}(x_{i,2})$ was computed as:

$$\frac{1}{2}\left(\hat{\phi}_1^{(2)}(x_{i,1}) + \hat{\phi}_1^{(2)}(x_{i,2})\right). \tag{33}$$

The proposed quadratic-mean-based fusions were computed as respectively:

$$\sqrt{\frac{1}{2}\left[\left(\hat{\phi}_1^{(1)}(x_{i,1})\right)^2 + \left(\hat{\phi}_1^{(1)}(x_{i,2})\right)^2\right]}. \tag{34}$$

and:

$$\sqrt{\frac{1}{2}\left[\left(\hat{\phi}_1^{(2)}(x_{i,1})\right)^2 + \left(\hat{\phi}_1^{(2)}(x_{i,2})\right)^2\right]}. \tag{35}$$

To evaluate and compare the two methods I examined the corresponding differential matching scores computed for the correct and incorrect information sources, that is:

$$\partial_1 = \hat{\phi}_1^{(1)}(x_{i,1}) - \hat{\phi}_1^{(2)}(x_{i,1}) = \|x_{i,1}\| - \|1 - x_{i,1}\|, \tag{36}$$

for the linear fusion, and

$$\partial_2 = \hat{\phi}_2^{(1)}(x_{i,2}) - \hat{\phi}_2^{(2)}(x_{i,2}) = \|x_{i,2}\| - \|1 - x_{i,2}\|. \tag{37}$$

for the proposed quadratic mean-based fusion where, without loss of generality, the correct prediction target for the first information source is taken to be 0 and for the second 1. It can be readily seen that a positive differential score corresponds to the correct source attribution (and a more confident decision for greater magnitudes thereof), a negative one to incorrect attribution (and a more mistaken confidence for greater magnitudes thereof), and a vanishing score to uncertain attribution.

3.2 Object recognition

My first case-study involves computer-based object recognition. This is an important problem in the spheres of computer vision and pattern recognition which has potential for use in a wide spectrum of practical applications. Examples range from motorway toll booths that automatically verify that the car type and its licence plate match the registration database, to online querying and

searching for an object of interest that was captured using a mobile phone camera. Object recognition has attracted much research attention [1, 2, 7, 8] and this interest has particularly intensified in recent years after significant advances towards practically viable systems have been made [20, 7, 8].

Here I consider fusion of texture-based and shape-based object descriptors. The former group has dominated research efforts to date [20] and has been highly successful in the matching of textured objects. However, it fails when applied on untextured objects (sometimes referred to as 'smooth') [7] Considering that their texture is not informative, characteristic discriminative information of smooth objects must be extracted from shape instead. Following the method described in [3] I extract and process the representations of the two modalities (texture and shape) independently. An object's texture is captured using a histogram computed over a vocabulary of textural words, learnt by clustering local texture descriptors extracted from the training data set. Similarly, a histogram over a vocabulary of elementary shapes, learnt by clustering local shape descriptors, is used to capture the object's shape. I adopt the standard SIFT descriptor as the basic building block of the texture representation and an analogous descriptor of local shape for the characterization of shape [4]. In both cases descriptors are matched using the Euclidean distance.

3.2.1 Results, analysis, and discussion

As explained in Section 3.1, the performances of the two fusion approaches was assessed by analysing the corresponding differential matching scores computed for the correct and incorrect information sources as summarized by (36) and (37).

My findings are summarized in Figure 1 which shows a plot of the cumulative density distribution of the relative improvement achieved by the proposed method (as before please note that the abscissa scale is logarithmic and that it is increasing in the leftward direction). Firstly let us make the remarkable observation that in *all* cases the proposed fusion strategy did at least as well as the simple weighted fusion. In approximately 66% of the cases the proposed quadratic fusion exhibited superior performance, performing on par in the remaining 34% of the cases. A further examination of the plot reveals an even stronger case for the proposed method – not only does it outperform simple weighted fusion in 66% of the cases but it does so with a great margin. For example in 30% of the cases the inter-class separation is increased 10-fold (i.e. by an order of magnitude).

3.3 Arrhythmia prediction

My second real-world case-study concerns automatic arrhythmia prediction from demographic data and physiological measurements. This is a challenging task of enormous importance in the provision of timely and informed health care provision to the population at risk of cardiac complications. This primarily includes individuals with preexisting cardiac problems or congenital factors, which are further modulated by various environmental factors such as high physical exertion [16] or the use of certain classes of drugs.

A normally functioning human heart maintains a remarkably well controlled heartbeat rhythm. Arrhythmias are abnormalities of this rhythm and are caused by physiological factors pertaining to electrical impulse generation or propagation. Arrhythmia types can be grouped under the umbrellas of three broad categories: tachycardias (overly high heartbeat rate), bradycardias (overly low heartbeat rate), and arrhythmias with an irregular heartbeat rate. While most arrhythmias are transient in nature and do not pose a serious health risk, some arrhythmias (particularly in high risk populations) can

have serious consequences such as stroke, cardiac arrest, or heart failure and death [18].

Arrhythmias can be diagnosed using nonspecific means, e.g. using a stethoscope or a tactile detection of pulse, or specific methods and in particular the electrocardiogram (ECG). Moreover, the rich information provided by the latter can be used to predict and detect the onset of arrhythmias by analysing electric impulse patterns [15]. In the present study I evaluate the proposed fusion methodology in the context of this prediction.

I used the dataset collected by Guvenir *et al.* [15] which is freely available[1]. The dataset comprises a rich set of demographic and anthropometric variables, such as each person's age, sex, height, and bodyweight, as well as a variety of features extracted from the person's ECG, including the heart rate and a series of characteristics of the corresponding signal deflections (the duration of the QRS interval, amplitudes of Q, R, and S waves etc); please refer to the original publication for full detail [15]. The target variable for the purpose of the present experiment can be considered to be binary valued, taking on the value 0 when arrhythmia is not present and 1 when it is; please see Table 1.

I consider two baseline classifiers. The first of these takes height, bodyweight, and the duration of the QRS interval as input variables (i.e. independent variables). The second one makes the prediction based on what are effectively integrals of Q, R, and S deflections (i.e. referred to as QRSA in the dataset description; please see the original reference for detail [15]), and the Q, R, S, and T deflections (i.e. referred to as QRSTA in the dataset description). Each classifier is built upon a generalized linear model (GLM) [19]. Recall that in a GLM the mean $E[Y]$ of the target, outcome variable Y is related to a linear predictor based on the independent variables through a link function $\mathcal{L}$:

$$E[Y] \equiv \mu = \mathcal{L}^{-1}(1 + \alpha x). \tag{38}$$

The variance $\text{Var}[Y]$ of Y is modelled as a function $\mathcal{V}$ of this mean:

$$\text{Var}[Y] = \mathcal{V}(\mu) = \mathcal{V}(\mathcal{L}^{-1}(1 + \alpha x)). \tag{39}$$

In my experiment the continuous output of a GLM is used to detect the presence of arrhythmia using what is effectively the nearest neighbour criterion: if the output is closer to 0 the prediction is taken to be negative, and if it is closer to 1 positive. The dataset contains 452 diagnosis cases of which 100 were used for training the classifiers; the remaining 352 were used for testing.

3.3.1 Results, analysis, and discussion

As in the previous experiment, the performances of the two fusion approaches was assessed by analysing the corresponding differential matching scores computed for the correct and incorrect information sources as summarized by (36) and (37).

My findings are summarized in Figure 2 which shows a plot of the cumulative density distribution of the relative improvement achieved by the proposed method (as before please note that the abscissa scale is logarithmic and that it is increasing in the leftward direction). We can again start with the observation which can be made by comparing the characteristics of the plot in Figure 2 with those of the plot in Figure 1 and observing that again the same qualitative behaviour is exhibited on this data too. In all cases the proposed fusion strategy did at least as well as the simple weighted fusion, with approximately 65% of the cases resulting in superior performance of the proposed quadratic fusion and the remaining 35% in on a par performance. As in the previous experiments when

---

[1]The dataset can be downloaded from http://archive.ics.uci.edu/ml/datasets/Arrhythmia.
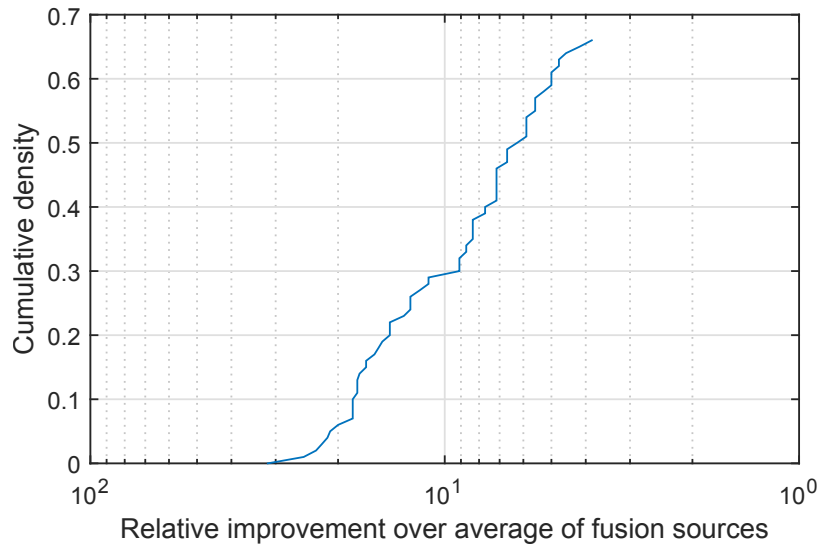
Figure 1: General object recognition: relative improvement in the correct information source identification of the proposed quadratic mean-based fusion over linear fusion, as the corresponding cumulative distribution function.

Table 1: An illustration of the adopted arrhythmia dataset, originally collected and described in detail by Guvenir *et al.* [15]. It comprises 279 input variables, of which only a small selection is shown here, and the corresponding target variable which for the purpose of the present experiment can be considered to be binary valued, taking on the value 0 when arrhythmia is not diagnosed and 1 when it is.

Age	Gender	Height	Body weight	QRS duration	...	Arrythmia
(years)		(cm)	(kg)	(ms)	...	(yes/no)
75	Male	190	80	91	...	Yes
56	Female	165	64	81	...	Yes
...	...	...	...	...	...	...
55	Male	175	94	100	...	No
...	...	...	...	...	...	

my method outperforms simple weighted fusion it does so with a significant margin with an over doubled separation increase in approximately 16% of the cases. Moreover I found that while simple weighted fusion failed to improve the prediction of the better performing baseline classifier in 15% of the cases, the same was the case with the proposed method in fewer than 7.8% of the cases.

3.4 Car accident fatality prediction

My third and final experiment herein concerns the inference of risk factors for car accident fatalities. In particular in this experiment I was interested in discovering what aspects of the context of an accident predict best if a fatality will occur. For this purpose I used the official statistics released by the government of the USA through the Fatality Analysis Reporting System (FARS) for the year 2011. These are freely publicly available[2]. The dataset comprises a number of person specific variables, such as age, sex,

---

[2]The dataset can be downloaded from http://www-fars.nhtsa.dot.gov/Main/index.aspx.

race, blood alcohol level, and drug use status, as well as a variety of variables pertaining to the context of the accident, including the type of the road and the state where the accident took place, and the weather conditions at the time (please see http://www-fars.nhtsa.dot.gov/Main/index.aspx for full description). The target variable for the purpose of the present experiment can be considered to be binary valued, taking on the value 1 when there has been a fatality and 0 otherwise; please see Table 2.

As in the previous experiment I consider two baseline classifiers, each built upon a generalized linear model. The first of these takes the person's age and blood alcohol level as input variables (i.e. independent variables). The second one makes the prediction based on atmospheric conditions and the road type. Following the same framework as described in the previous section, the continuous output of a GLM is used to predict the occurrence of a fatality using what is effectively the nearest neighbour criterion: if the output is closer to 0 the prediction is taken to be negative, and if it is closer to 1 positive. The dataset contains 5000 accidents of which 1000 ran-
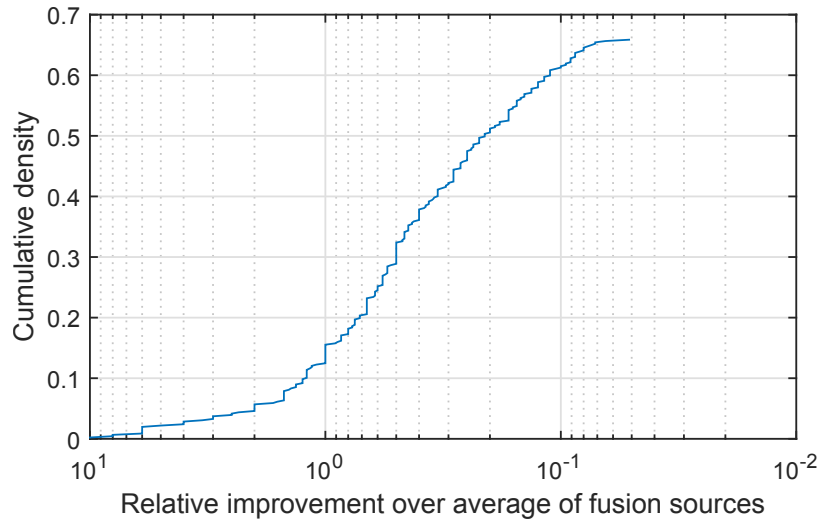
Figure 2: Arrhythmia prediction: relative improvement in the correct information source identification of the proposed quadratic mean-based fusion over linear fusion, as the corresponding cumulative distribution function.

Table 2: An illustration of the adopted car accidents dataset released by the government of the USA through the Fatality Analysis Reporting System (FARS) for the year 2011. It comprises 15 input variables, of which only a selection is shown here, and the corresponding target variable which for the purpose of the present experiment can be considered to be binary valued, taking on the value 1 when there has been a fatality and 0 when not.

Clear atmospheric conditions (yes/no)	Driver age	Alcohol blood level	Driver gender	...	Fatalities (yes/no)
Yes	27	0	Male	...	No
Yes	60	0	Female	...	Yes
...	...	...	...	...	...
No	20	0.21	Male		Yes
...	...	...	...	...	...

domly selected cases were used for training the baseline classifiers with the remainder employed for performance evaluation.

3.4.1 Results, analysis, and discussion

I followed the same evaluation methodology as in the preceding experiments. In the same vein I display my findings in Figure 3 which shows a plot of the cumulative density distribution of the relative improvement achieved by the proposed method (as before please note that the abscissa scale is logarithmic and that it is increasing in the leftward direction). Yet again, in agreement with the results obtained in all experiments I conducted, we can observe the same functional characteristics in the plot of Figure 3 and of those in Figures 1 and 2. In further agreement with the previous experiments is my finding that in all instances of car accidents used for evaluation, the proposed fusion strategy did at least as well as the simple weighted fusion in terms of its predictive ability. With regard to this point, it is interesting to note that notwithstanding the remarkable qualitative similarity of the CDFs corresponding to different experiments, there are some quantitative differences.

For example, note that in the present experiment approximately 73% of the cases resulted in superior performance of the proposed quadratic fusion and the remaining 27% in on a par performance. The proportion of evaluation instances yielding superior performance is thus higher than e.g. in the object recognition experiment described in Section 3.2. However, the resulting benefit is smaller. For example, while in the object recognition experiment about 30% of the evaluation cases result in at least 10-fold class separation increase, in the present experiment in no case is the benefit as large. This is most likely a consequence of the inherent information content in the data itself, rather than of some algorithmic aspect of the proposed method. In other words, the lesser advantage (though still consistent and observed in an overwhelming number of cases) of using the proposed method stems from greater redundancy across the fused information sources.

4. REFERENCES

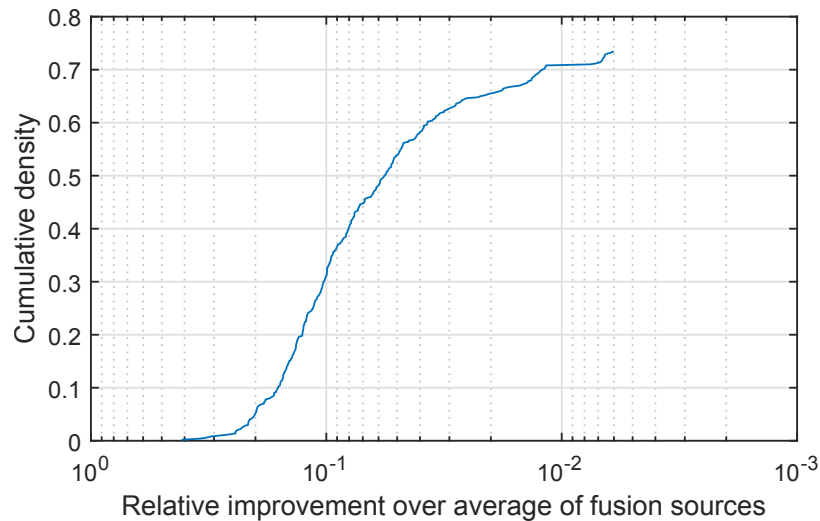[1] G. Aggarwal and D. Roth. Learning a sparse representation

Figure 3: Fatalities in car accidents prediction: relative improvement in the correct information source identification of the proposed quadratic mean-based fusion over linear fusion, as the corresponding cumulative distribution function.

for object detection. *In Proc. European Conference on Computer Vision*, 2002.

[2] A. Ahmadyfard and J. Kittler. A comparative study of two object recognition methods. *In Proc. British Machine Vision Conference*, pages 1–10, 2002.

[3] O. Arandjelović. Matching objects across the textured–smooth continuum. *In Proc. Australasian Conference on Robotics and Automation*, pages 354–361, 2012.

[4] O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85.

[5] O. Arandjelović and R. Cipolla. A new look at filtering techniques for illumination invariance in automatic face recognition. *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 449–454, 2006.

[6] O. Arandjelović, R. I. Hammoud, and R. Cipolla. Thermal and reflectance based personal identification methodology in challenging variable illuminations. *Pattern Recognition*, 43(5):1801–1813, 2010.

[7] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. *In Proc. IEEE International Conference on Computer Vision*, pages 375–382, 2011.

[8] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, USA, 2007.

[10] R. S. Ghiass and O. Arandjelović. Highly accurate gaze estimation using a consumer RGB-depth sensor. *In Proc. International Joint Conference on Artificial Intelligence*, pages 3368–3374, 2016.

[11] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague. Infrared face recognition: a literature review. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 2791–2800, 2013.

[12] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague. Infrared face recognition: a comprehensive review of methodologies and databases. *Pattern Recognition*, 47(9):2807–2824, 2014.

[13] V. Ginsburgh, M. Monzak, and A. Monzak. Red wines of médoc: What is wine tasting worth? *Journal of Wine Economics*, 8(2):159–188, 2013.

[14] Y. Guan, X. Wei, C. T. Li, and Y. Keller. People identification and tracking through fusion of facial and gait features. *Biometric Authentication*, pages 209–221, 2014.

[15] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin. A supervised machine learning algorithm for arrhythmia analysis. *In Proc. Computers in Cardiology Conference*, pages 433–436, 1997.

[16] C. Lengyel, A. Orosz, P. Hegyi, Z. Komka, A. Udvardy, E. Bosnyák, E. Trájer, G. Pavlik, M. Tóth, T. Wittmann, J. G. Papp, A. Varró, and I. Baczkó. Increased short-term variability of the qt interval in professional soccer players: Possible implications for arrhythmia prediction. *PLOS ONE*, 6(4):e18751, 2011.

[17] A. Mackinnon and R. Mulligan. Combining cognitive testing and informant report to increase accuracy in screening for dementia. *The American Journal of Psychiatry*, 155:1529–1535, 1998.

[18] R. J. Myerburg, A. J. Interian, R. M. Mitrani, K. M. Kessler, and A. Castellanos. Frequency of sudden cardiac death and profiles of risk. *The American Journal of Cardiology*, 80(5):10F–19F, 1997. (supplement 2).

[19] J. Nelder and R. Wedderburn. Generalized linear models. *Proceedings of the Royal Society, Series A*, 135(3):370–384, 1972.

[20] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *In Proc. IEEE International Conference on Computer Vision*, pages 370–377, 2005.