# Online Learning for Multimodal Data Fusion With Application to Object Recognition

Shahin Shahrampour, Mohammad Noshad, Jie Ding, and Vahid Tarokh

***Abstract*—We consider online multimodal data fusion, where the goal is to combine information from multiple modes to identify an element in a large dictionary. We address this problem in the context of object recognition by focusing on tactile sensing as one of the modes. Using a tactile glove with seven sensors, various individuals grasp different objects to obtain 7-D time series, where each component represents the pressure sequence applied to one sensor. The pressure data of all objects is stored in a dictionary as a reference. The objective is to match a streaming vector time series from grasping an unknown object to a dictionary object. We propose an algorithm that may start with prior knowledge provided by other modes. Receiving pressure data sequentially, the algorithm uses a dissimilarity metric to modify the prior and form a probability distribution over the dictionary. When the dictionary objects are dissimilar in shape, we empirically show that our algorithm recognize the unknown object even with a uniform prior. If there exists a similar object to the unknown object in the dictionary, our algorithm needs the prior from other modes to detect the unknown object. Notably, our algorithm maintains a similar performance to standard offline classification techniques, such as support vector machine, with a significantly lower computational time.**

***Index Terms*—Online learning, mirror descent, tactile sensing, object recognition.**

## I. Introduction

ANALYZING a system or task requires information from various measurements, experiments, and multiple resources, in that natural phenomena often have complex characteristics [1]. Due to advancements in hardware technology, availability of high-resolution sensors, and abundance of large-scale sensory data, one can employ numerous resources to analyze a complex system. These resources provide "multimodal" data that can be fused to accomplish a task without recourse to a single dataset that is unreliable. Fusing multiple datasets has roots in statistics dating back several decades [2], [3] and has been employed in various applications (e.g., spatial-temporal prediction [4], or estimation of the state-space projection operator [5]).

Manuscript received August 4, 2017; accepted September 13, 2017. Date of publication September 18, 2017; date of current version August 28, 2018. This work was supported by DARPA under Grant N66001-15-C-4028 and Grant W911NF-14-1-0508. This brief was recommended by Associate Editor L.-P. Chau. *(Corresponding author: Shahin Shahrampour.)*

S. Shahrampour, J. Ding, and V. Tarokh are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: shahin@seas.harvard.edu; jieding@g.harvard.edu; vahid@seas.harvard.edu).

M. Noshad is with VLNComm, Charlottesville, VA 22911 USA (e-mail: mnoshad@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2017.2754141

Motivated by advantages of multimodal data analysis, we consider online fusion of multimodal data for identifying an element in a large dictionary. We present this brief in the context of object recognition, where the goal is to detect an unknown object among a finite number of objects. Various object recognition methods use image inputs to detect an object [6], and many machine learning algorithms such as Support Vector Machine (SVM), neural networks, and boosting are employed for object recognition [7]–[9]. These methods perform quite well on large datasets, but incur a prohibitive computational cost especially in the training phase. The natural limitations of these methods have stimulated interest in multimodal object detection [10]–[12]. In this context, a natural mode to embed in a learning system is the "sense of touch", since humans naturally learn about objects through grasping and dextrous manipulation. Tactile sensing provides useful information about objects (e.g., shape, size, and material), and humans are excellent at identifying objects solely based on touch [13].

In this brief, we study multimodal object recognition on humans, which can serve as a good starting point for studying robots. Focusing on tactile sensing as one of the modes, we use a tactile glove with seven sensors to grasp different objects and obtain a 7-dimensional vector time series, each component of which represents the pressure applied to one sensor. We then construct a dictionary using the vector time series corresponding to various objects grasped by different individuals. Given the dictionary, the problem is to match streaming samples from grasping an unknown object to an object in the dictionary. The key observation is that "proper" grasps most likely have similar patterns, though there might be a scaling in the pressures experienced by the hand from time to time. For instance, this is due to person A generally grasping objects more firmly than person B, or person A experiencing more hand tremors compared to person B. Therefore, to measure the dissimilarity of the unknown object compared with dictionary objects, we introduce a metric (loss function) that is invariant with respect to scaling of the dictionary objects.

We propose an algorithm, referred to as Optimistic Exponential Weights (`OEW`), which builds on the Optimistic Mirror Descent (`OMD`) algorithm developed in [14]. As shown in Fig. 1, our algorithm may receive *prior* information about the unknown object from other "modes" or object recognition techniques (e.g., SVM, boosting, and etc.) and incorporates the given prior knowledge with the obtained grasp data to identify the unknown object. The `OEW` algorithm sequentially receives the stream of *input* pressure data obtained by grasping the *unknown* object. The algorithm uses the pressure data to iteratively calculate a loss function, which measures the similarity between the unknown object and dictionary objects. Using the loss function, the algorithm forms a probability distribution (belief) over the dictionary objects, giving each object

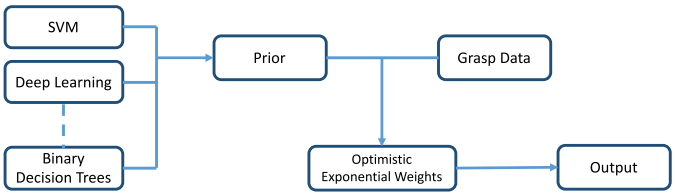


Fig. 1. A diagram of the multimodal learning: OEW combines the grasp data with prior information to infer the unknown object. The prior can be the output of another object recognition algorithm.

a weight. Then, the algorithm refines its previous belief using a *correction* step.

We implement our algorithm using the dataset collected by the tactile glove and observe a promising empirical performance. When the shape and grasping pattern of the unknown object are different enough from those of the dictionary objects, our method outperforms the Exponential Weights in identifying the unknown object even with a uniform prior. However, when the unknown object is similar in shape to another object in the dictionary, our method fails to uniquely distinguish the unknown object, though it can rule out the possibility of dissimilar objects being the unknown object. In such cases, we can incorporate the prior information gained from other recognition methods to uniquely identify the unknown object. In addition, the performance of our algorithm is comparable to standard offline classification techniques (e.g., SVM and Logistic Regression) with a significantly lower computational cost.

*Related work on tactile sensing for object recognition:* There exists a large body of literature on tactile sensing for a variety of tasks. We mention some of the recent works here, and due to space limitation, we refer the reader to surveys for a fair treatment of the subject (see [15]–[18]). In the context of haptic object recognition, Pezzementi *et al.* [19] interpret sensor readings as tactile images to explore the unknown object and reconstruct its appearance. Their work is closely connected to that of Schneider *et al.* [20], where "bag-of-features" is used for object identification. Other works have addressed object recognition based on extracting key haptic features [21], Bayesian estimation of position and orientation [22], developing extreme kernel sparse learning [23], fusing tactile and image data [24], active touch [25], and manifold learning [26]. In contrast to the existing literature, our work combines *online learning* and *tactile sensing* for object recognition. This approach is useful for on-the-fly processing of data, where standard offline methods suffer from prohibitive computational cost.

*Related work on online learning:* Online learning has been intensely studied in the literature of machine learning [27] and has inspired many algorithms for engineering applications (e.g., online independent component analysis [28] and approximate/adaptive dynamic programming [29]). One of the classical techniques in online learning is prediction with expert advice [27], which has been successfully applied to many problems including online clustering and classification [30], [31]. This brief is particularly related to online optimization methods that take advantage of the similarity in adversarial sequences [14], [32]. Unlike traditional schemes, these algorithms work based on a couple of updates per iteration. Subsequently, the performance captures the similarity of the loss sequences over time, interpolating between *adversarial* and *non-adversarial* environments.

*Notation:* For each integer $K$ we define $[K] \coloneqq \{1, \ldots, K\}$ to represent the set of integers smaller than or equal to $K$. We represent the $k$-th element of a vector $x$ by $x(k)$, and denote the $\mathcal{L}_p$-norm operator by $\|\cdot\|_p$. The inner product of vectors $x$ and $y$ is denoted by $<x, y>$.

## II. Problem Statement

In this section, we describe the object recognition problem in an online learning scheme, discuss the properties of tactile data, and propose our methodology. We provide our results in Section III, where we discuss our data collection and numerical experiments.

Consider a tactile glove with several sensors that can measure the pressure applied to them. Using such a glove, we can build a database (dictionary) consisting of pressure samples by grasping various objects. Given the dictionary, the problem is to associate a stream of input samples, obtained by grasping an *unknown* object, to a single object in the dictionary. The key to accomplish this task is to define a suitable *loss* function to measure the similarity of input samples with those of the dictionary. We will elaborate on the choice of the loss function in Section II-A.

It is natural to assume that the measured samples are corrupted with noise. The noise can be either *stochastic* or *adversarial*. In the stochastic model, the samples are assumed to be corrupted by a noise with a fixed statistical distribution, i.e., the noise realizations are independent and identically distributed. On the other hand, adversarial analysis often provides worst-case performance guarantees without any assumption on the distribution of the noise. We restrict our attention to the adversarial case, motivated by the fact that even an additive noise to the samples might result in a complicated distribution on the loss function (as we shall see in Section II-A).

In mathematical representation, we assume the number of objects in the dictionary to be $K$, i.e., $\mathcal{D} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, each $\mathcal{C}_i$ denoting an object for $i \in [K]$. The entire dataset $\{x_s\}_{s=1}^{MT}$ consists of $MT$ samples arriving as follows: at time $t \in [T]$, $M$ samples $\{x_s\}_{s=M(t-1)+1}^{Mt}$ are observed, where each $x_s$ is a vector. The algorithm compares the samples with the corresponding $M$ samples of object $k$ in the dictionary, i.e., with $\{y_{s,k}\}_{s=M(t-1)+1}^{Mt}$ for $k \in [K]$. It then calculates the loss $\ell_t(k)$, for each $k \in [K]$. The loss function serves as a dissimilarity measure, quantifying the similarity among the objects. According to the loss function, the algorithm forms a probability distribution $p_t$ over the dictionary space. The algorithm will be further explained in Section II-B.

Note that the tactile data satisfies two properties: (i) The samples of the unknown object are temporally aligned with those of the objects in the dictionary. That is, there is no delay or shift in the input signal, and there are no missed samples. (ii) The samples of the unknown object may be scaled versions of the true object in the dictionary. The scale transition in samples satisfies a *smoothness* property in the sense that samples closer to each other are scaled in a similar fashion. In other words, the scaling does not change drastically from one instance to another instance; there is some regularity in the signal.

The first assumption is important to our analysis since existence of shifted (or missed) input samples requires more considerations in the introduction of loss function. The intuition behind the second assumption on the grasp data is that a person generally does not switch from holding an object firmly

or loosely for no reason. Therefore, the pressure changes smoothly during a stable, uninterrupted grasp of an object. We now need to introduce a proper loss function for the algorithm.

### A. Scale-invariant Loss Function

The loss function characterizes the dissimilarity among objects. We need to choose a loss function that captures the scale-invariance property. The data arrives sequentially in batches of $M$ samples. That is, the $t$-th group of $M$ samples are those belonging to the interval $\mathcal{I}_t := [M(t-1)+1, Mt]$ for $t \in [T]$. Recall that any sample $x_s$ corresponds to $y_{s,k}$ in object $k$ for $k \in [K]$. We now define the following loss function for any object index $k \in [K]$

$$\ell_t(k) := \min_\alpha \sum_{s\in\mathcal{I}_t} \left\|x_s - \alpha y_{s,k}\right\|_2^2 = \sum_{s\in\mathcal{I}_t} \left\|x_s\right\|_2^2 - \frac{\left|\sum\limits_{s\in\mathcal{I}_t} \langle x_s, y_{s,k}\rangle\right|^2}{\sum\limits_{s\in\mathcal{I}_t} \left\|y_{s,k}\right\|_2^2}, \tag{1}$$

which compares the $t$-th batch of samples with their corresponding samples in the $k$-th object. Note that this choice of loss function removes the local scaling of dictionary elements, i.e., when $x_s = \alpha y_{s,k}$ for some $\alpha > 0$ over the interval $\mathcal{I}_t$, we have that $\ell_t(k) = 0$. Without loss of generality, we assume that the loss function is bounded by unity. This can be achieved simply by dividing the vector $\ell_t$ by $\|\ell_t\|_\infty$ for each $t \in [T]$.

### B. Algorithm

Let us now describe Optimistic Exponential Weights (OEW) inspired by Optimistic Mirror Descent (OMD) developed in [14]. The OEW algorithm works based on a probabilisitic viewpoint. The algorithm is initialized with a prior $p'_0$ provided by an external strategy as in Fig. 1. Then, it starts modifying the prior using the sequentially-fed grasp data as follows: at time $t$, it predicts the unknown object by forming a probability distribution $p_t$ over the dictionary space; then, it refines its prediction with another probability distribution $p'_t$ before making a new prediction $p_{t+1}$. The refinement step is what makes the algorithm "optimistic". The description of OEW is given in Algorithm 1. We can rewrite the algorithm in the form of the following updates

$$p'_t(k) = \frac{p'_{t-1}(k)e^{-\eta_t \ell_t(k)}}{\left\langle p'_{t-1}, e^{-\eta_t \ell_t}\right\rangle}, \quad p_t(k) = \frac{p'_{t-1}(k)e^{-\eta_t \ell_{t-1}(k)}}{\left\langle p'_{t-1}, e^{-\eta_t \ell_{t-1}}\right\rangle}, \tag{2}$$

for each $k \in [K]$, with the choice of step size $\eta_t = \min\{1, (\sqrt{D_{t-1}} + \sqrt{D_t})^{-1}\}$, where $D_t = \sum_{s=1}^{t-1} \|\ell_s - \ell_{s-1}\|_\infty^2$. It has been proved theoretically in [14] that the OMD algorithm has good performance using the prediction step as well as the adaptive step-size. Since our algorithm can be viewed as a special case of the OMD algorithm, it is immediate that the theoretical guarantees hold true. In this brief, we are interested in examining its performance in practice, when the algorithm is employed for an object recognition task.

We remark that the computational complexity of OEW consists of two parts. The online procedure requires $\mathcal{O}(MTK)$ computations, which mainly involves calculation of the loss function in (1). Additionally, we have the prior $p'_0$ provided by other algorithms. Denoting the computational cost of the prior by $C(p'_0)$, the overall complexity would be $\mathcal{O}(MTK + C(p'_0))$. The cost $C(p'_0)$ completely depends on the numerical solution used to obtain $p'_0$. Note that if we want to analyze the pressure signals using standard (offline) SVM, we at least have a computational complexity of $\mathcal{O}((MTK)^2)$ which is much larger than our algorithm (unless $C(p'_0)$ is large).

**Algorithm 1** Optimistic Exponential Weights

Initialize : Let $p'_0$ be a prior distribution on $[K]$. $D_1 = 0$, $\ell_0 = 0$, $\eta_1 = 1$.
**for** $t = 1$ to $T$ **do**
  $p_t(k) \propto p'_{t-1}(k)e^{-\eta_t \ell_{t-1}(k)}$ for all $k \in [K]$
  Receive $M$ samples and calculate the loss $\ell_t$
  $p'_t(k) \propto p'_{t-1}(k)e^{-\eta_t \ell_t(k)}$ for all $k \in [K]$
  $D_{t+1} = D_t + \|\ell_t - \ell_{t-1}\|_\infty^2$
  $\eta_{t+1} = \min\left\{1, \left(\sqrt{D_{t+1}} + \sqrt{D_t}\right)^{-1}\right\}$
**end for**

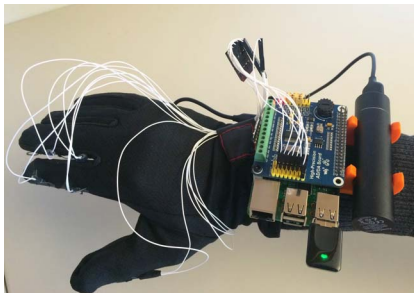

Fig. 2. The tactile glove used in our experiments has seven sensors (7 degrees of freedom) connected to a Raspberry Pi. Once an object is grasped, the sensors measure the pressure and the Raspberry Pi reads the data.

## III. Experiments

We now evaluate the performance of the OEW algorithm. Our focus is on tactile sensing (as one of the modes) for the experiment design, assuming that the prior encapsulates the information provided by other modes. As mentioned in Section II, the tactile glove is equipped with seven sensors, each of which measures the pressure applied to it. The algorithm has access to a pre-defined, offline dictionary consisting of pressure data for various objects. In other words, there exists a reference of "good grasp" for each of the objects. Initially, the algorithm receives a prior over the dictionary space, which is provided by an external recognition algorithm. It then sequentially refines this distribution after receiving a stream of new samples.

The parameter $M$ for the loss function (1) should be fixed. We use $M \in \{10, 20\}$ for our experiment. A small $M$ (e.g., $M \leq 5$) can make the prediction unrealistic as it allows too much scale variation along the samples. On the other end of the spectrum, large $M$ would restrict the scaling freedom.

### A. Tactile Glove & Data Collection

We collect the pressure data using the tactile glove shown in Fig. 2. The sensors are placed on the fingers and the palm of the glove where the contacts are most likely to occur. The sensors are connected to 8-channel 24-bit high-precision analog-to-digital converter (ADC), and a Raspberry Pi is used to read the output of the ADC. The pressure on the sensors is sampled at 250 Hz.

We build the dictionary using 15 objects shown in Fig. 3. Several individuals grasp each object *properly* (in a stable manner) for 4 seconds and record the pressure applied to each sensor. The resulting sequence for each object grasped by each individual consists of 1000 7-dimensional samples. In turn, the

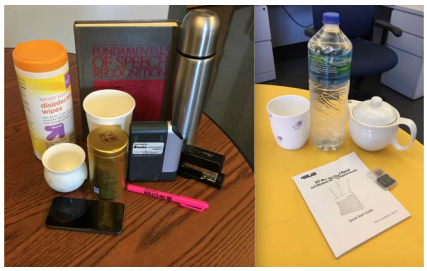

Fig. 3. The objects used in the dictionary: we hold each object for a few seconds (stable grasp) and record the samples.

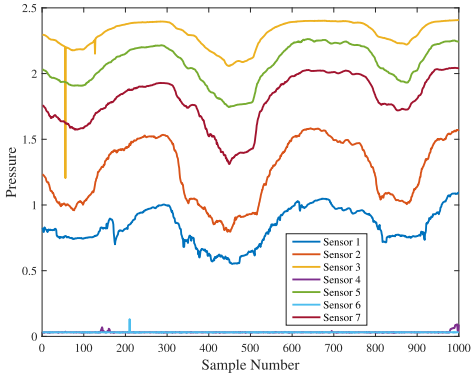


Fig. 4. The input signal consisting of 1000 samples arriving in a sequential fashion. The signal is recorded when holding the book for approximately four seconds. The non-monotonicity indicates that the grasp can become loose or firm along the way.

dictionary comprises all these sequences. Ideally, the dictionary should be a large database, but in this brief we only show a proof of concept with 15 objects.

For the next step, another person is asked to grasp the unknown object and to record its corresponding signal. For instance, for the experiment considered in the next section, the unknown object is the book shown in Fig. 3. 1000 samples of its corresponding signal, represented in Fig. 4, are sequentially fed to the OEW algorithm as input. With $M = 20$, having 1000 samples means that at each iteration the algorithm receives 20 samples, and therefore operates for 50 iterations. The algorithm compares the samples with their corresponding samples in the dictionary and determines which object maximizes $p_t$.

### B. Performance With No Similar Objects

In this section, we evaluate the performance of the OEW algorithm when the unknown object is not similar to dictionary objects. As we see in the Fig. 3, the book has a different shape compared to other objects in the dictionary, so we consider it to be the unknown object. In this case, we assume a uniform prior over the dictionary and expect that $p_t$ will converge to a delta distribution centered on the correct object (book). In fact, in Fig. 5, we observe that not only does the OEW algorithm converge, but also it outperforms Exponential Weights in identifying the book. The probability distribution $p_t$ converges to a delta distribution concentrated on the true object. As we mentioned before, unlike the classical Exponential Weights, the OEW algorithm features a correction step, which results in better performance.

We further compare the performance and computational cost of our algorithm with three standard offline classification methods: linear SVM (L-SVM), kernel SVM (K-SVM), and

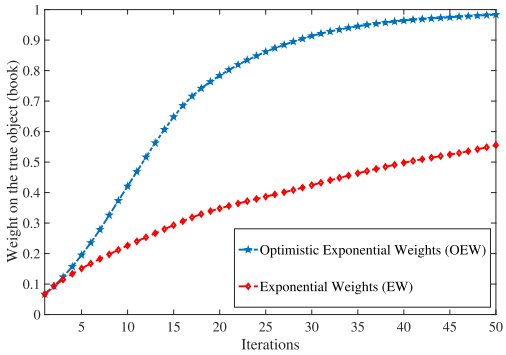


Fig. 5. The performance of the OEW algorithm surpasses that of the classical Exponential Weights algorithm. The algorithm weights the book almost with 1 (complete identification) after 50 rounds of iterations.

TABLE I
THE PERFORMANCE AND TIME COST OF OFFLINE CLASSIFICATION ALGORITHMS VERSUS OUR ALGORITHM (OEW)

| Algorithm | L-SVM | K-SVM | LR | OEW |
|---|---|---|---|---|
| Probability | 0.92 | 1 | 0.999 | 0.9834 |
| Time (sec) | 0.778 | 8.83 | 1.52 | 0.128 |

logistic regression (LR). For K-SVM, we use the RFB (Gaussian) kernel with the best parameter over [0.001, 1000]. These methods are provided with the dictionary as training samples and the whole input of our algorithm as test samples. In order to compare the output of these methods with our algorithm (whose output is a probability vector), we perform the following: for each method, we run a binary classification on one object versus other objects and find the ratio of predicted labels corresponding to that object to the whole predicted labels (1000). We then normalize these ratios such that their sum is equal to one. Table I tabulates the probability assigned to the book as well as the computation time of each algorithm.

All experiments are performed on a machine operating at 2.9GHz with 16GB RAM. As we can see, OEW outperforms L-SVM with a less computational time. Compared to LR and K-SVM, OEW maintains a similar performance with 0.084 and 0.014 computational-cost ratio, respectively.

### C. Performance in the Presence of Similar Objects

When the unknown object is similar in shape to another object in the dictionary, the prior plays a key role in the object identification. Tactile sensing provides information about the sense of touch, and therefore, when two objects are similar in shape and weight, an algorithm that relies solely on tactile sensing would not be completely successful. In such cases, the algorithm needs the prior from another algorithm to uniquely identify the unknown object.

To illustrate this, let us consider another experiment. We record a set of pressure samples corresponding to grasping a water bottle which is similar in shape and mass to the flask shown in Fig. 3. We then ask another person to provide another set of samples for the water bottle as the unknown object. First, we run the OEW algorithm with a uniform prior for 50 rounds ($M = 20$) and call it U-OEW. Next, we the assign a non-uniform prior (with more weight towards bottle) and run the algorithm with the new prior (NU-OEW). We also run SVM for multi-class classification with linear (L-SVM), polynomial

TABLE II
THE PERFORMANCE AND TIME COST OF OFFLINE CLASSIFICATION ALGORITHMS VERSUS OUR ALGORITHM WITH A UNIFORM PRIOR (U-OEW) AND A NON-UNIFORM PRIOR (NU-OEW)

| Algorithm | L-SVM | P-SVM | G-SVM | U-OEW | NU-OEW |
|---|---|---|---|---|---|
| Bottle Probability | 0.446 | 0.438 | 0.737 | 0.5481 | 0.7844 |
| Flask Probability | 0.554 | 0.562 | 0.263 | 0.4371 | 0.2085 |
| Others Probability | 0 | 0 | 0 | 0.0149 | 0.0071 |
| Time (sec) | 2.06 | 2.09 | 34.12 | 0.128 | N/A |

(P-SVM), and Gaussian kernel (G-SVM). The output of these algorithm is reported in Table II.

All algorithms are able to rule out all the objects in Fig. 3 except for the water flask which is similar to the water bottle. Each algorithm assigns a fair amount of probability to the flask. Compared to U-OEW, the best offline technique is G-SVM which is the best outcome over various parameters for the kernel; however, that comes at the cost of almost 266.5 more computational time. In addition, our algorithm with a non-uniform prior (NU-OEW) can slightly outperform G-SVM. The computational cost of NU-OEW depends on how we obtain the prior from another modes. In summary, using our online algorithm, we can maintain a similar performance to offline techniques while drastically decreasing the computational time.

## IV. CONCLUSION

In this brief, we considered object recognition via tactile sensing, where the goal is to identify an unknown object that belongs to a known dictionary. We proposed an online algorithm that is a variant of Exponential Weights and exhibits promising practical guarantees. In particular, when applied to tactile data, it outperforms Exponential Weights in identifying the unknown object. This work opens several avenues for future works; in particular, we plan to (i) study the performance of our algorithm with a more complete dictionary, and (ii) deal with shifted and missing samples. In this brief, we compare equal number of samples from an unknown object to dictionary objects. It would be interesting to propose a novel dissimilarity metric to quantify the similarity of unequal number of samples.

## ACKNOWLEDGMENT

The authors would like to thank Professor Zhu's group at UCLA Center for Vision, Cognition, Learning and Autonomy for providing the tactile glove and Raspberry Pi for this project. They would also thank Ahmad Beirami, Kathryn Heal, and the anonymous reviewers for their helpful comments on this brief.

## REFERENCES

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.
[2] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[3] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
[4] Y. Xia, H. Leung, and H. Chan, "A prediction fusion method for reconstructing spatial temporal dynamics using support vector machines," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 1, pp. 62–66, Jan. 2006.
[5] P. Cui, Q. Pan, K. Zhao, G. Wang, and J. Li, "Estimation of the projection operator in a multiresolution multisensor data fusion scheme," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 12, pp. 1343–1347, Dec. 2006.
[6] R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision*, vol. 5. New York, NY, USA: McGraw-Hill, 1995.
[7] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2155–2162.
[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[9] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 854–869, May 2007.
[10] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller, "Integrating visual and range data for robotic object detection," in *Proc. Workshop Multi Camera Multi Modal Sensor Fusion Algorithms Appl. (M2SFA2)*, 2008.
[11] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Hamburg, Germany, 2015, pp. 681–687.
[12] M. Villamizar, A. Garrell, A. Sanfeliu, and F. Moreno-Noguer, "Random clustering ferns for multimodal object recognition," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2445–2460, 2017.
[13] R. L. Klatzky, S. J. Lederman, and V. A. Metzger, "Identifying objects by touch: An 'expert system,"' *Perception Psychophys.*, vol. 37, no. 4, pp. 299–302, 1985.
[14] S. Rakhlin and K. Sridharan, "Optimization, learning, and games with predictable sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3066–3074.
[15] R. D. Howe, "Tactile sensing and control of robotic manipulation," *Adv. Robot.*, vol. 8, no. 3, pp. 245–261, 1993.
[16] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—From humans to humanoids," *IEEE Trans. Robot.*, vol. 26, no. 1, pp. 1–20, Feb. 2010.
[17] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—A review," *Sensors Actuators A Phys.*, vol. 167, no. 2, pp. 171–187, 2011.
[18] B. D. Argall and A. G. Billard, "A survey of tactile human–robot interactions," *Robot. Auton. Syst.*, vol. 58, no. 10, pp. 1159–1176, 2010.
[19] Z. Pezzementi, E. Plaku, C. Reyda, and G. D. Hager, "Tactile-object recognition from appearance information," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 473–487, Jun. 2011.
[20] A. Schneider *et al.*, "Object identification with tactile sensors using bag-of-features," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, St. Louis, MO, USA, 2009, pp. 243–248.
[21] N. Gorges, S. E. Navarro, D. Göger, and H. Wörn, "Haptic object recognition using passive joints and haptic key features," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Anchorage, AK, USA, 2010, pp. 2349–2355.
[22] A. Petrovskaya, O. Khatib, S. Thrun, and A. Y. Ng, "Bayesian estimation for autonomous object manipulation based on tactile sensors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Orlando, FL, USA, 2006, pp. 707–714.
[23] H. Liu, J. Qin, F. Sun, and D. Guo, "Extreme kernel sparse learning for tactile object recognition," *IEEE Trans. Cybern.*, to be published.
[24] J. Yang, H. Liu, F. Sun, and M. Gao, "Object recognition using tactile and image information," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Zhuhai, China, 2015, pp. 1746–1751.
[25] N. F. Lepora, U. Martinez-Hernandez, and T. J. Prescott, "Active touch for robust perception under position uncertainty," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Karlsruhe, Germany, 2013, pp. 3020–3025.
[26] D. Tanaka, T. Matsubara, K. Ichien, and K. Sugimoto, "Object manifold learning with action features for active tactile object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Chicago, IL, USA, 2014, pp. 608–614.
[27] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[28] C.-L. Chang, K.-W. Fan, I.-F. Chung, and C.-T. Lin, "A recurrent fuzzy coupled cellular neural network system with automatic structure and template learning," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 8, pp. 602–606, Aug. 2006.
[29] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 10, pp. 693–697, Oct. 2012.
[30] V. Vovk and F. Zhdanov, "Prediction with expert advice for the brier game," *J. Mach. Learn. Res.*, vol. 10, pp. 2445–2471, Dec. 2009.
[31] A. Choromanska and C. Monteleoni, "Online clustering with experts," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2012, pp. 227–235.
[32] C.-K. Chiang *et al.*, "Online optimization with gradual variations," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–20.