BANDIT FRAMEWORK FOR SYSTEMATIC LEARNING IN WIRELESS VIDEO-BASED FACE RECOGNITION

Onur Atan, Cem Tekin, Mihaela van der Schaar

University of California, Los Angeles, CA
Dept. of Electrical Engineering

Yiannis Andreopoulos

University College London, UK, Dept. of
Electronic and Electrical Engineering

ABSTRACT

In most video-based object or face recognition services on mobile devices, each device captures and transmits video frames over wireless to a remote computing service (a.k.a. "cloud") that performs the heavy-duty video feature extraction and recognition tasks for a large number of mobile devices. The major challenges of such scenarios stem from the highly-varying contention levels in the wireless local area network (WLAN), as well as the variation in the task-scheduling congestion in the cloud. In order for each device to maximize its object or face recognition rate under such contention and congestion variability, we propose a systematic learning framework based on *multi-armed bandits*. Unlike well-known reinforcement learning techniques that exhibit very slow convergence rates when operating in highly-dynamic environments, the proposed bandit-based systematic learning quickly approaches the optimal transmission and processing-complexity policies based on feedback on the experienced dynamics (contention and congestion levels). Comparisons against state-of-the-art reinforcement learning methods demonstrate that this makes our proposal especially suitable for the highly-dynamic levels of wireless contention and cloud scheduling congestion.

Index Terms— multi-armed bandits, learning, face recognition, cloud computing, wireless contention, scheduling congestion

1. INTRODUCTION

Many of the envisaged applications and services for wearable sensors, smartphones, tablets or portable computers in the next ten years will involve analysis of video streams for event, action, object or user recognition [1, 2]. In this process, they experience time-varying channel conditions, traffic loads and processing constraints at the remote cloud-computing servers where the data analysis takes place. Examples of early commercial services in this domain include Google Goggles, Google Glass, Facebook automatic face tagging [3] and Microsoft's Photo Gallery face recognition.

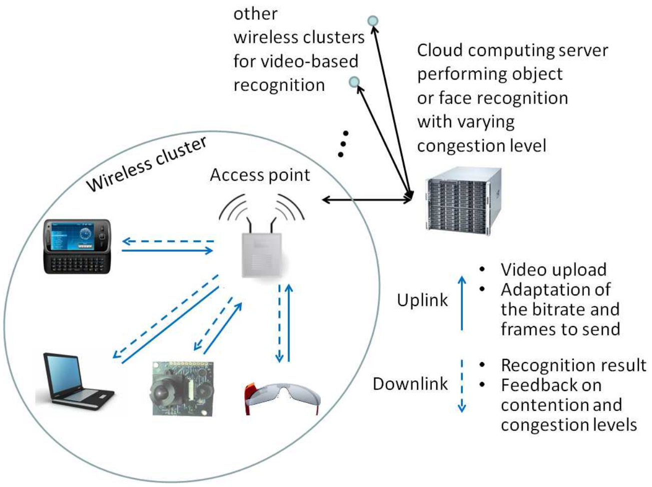

Fig. 1. Illustration of object or face recognition via adaptive wireless video transport to a remote computing server.

Figure 1 presents an example of such deployments. Video content producers include several types of sensors, mobile phones, as well as other low-end portable devices, that capture, encode (typically via a hardware-supported MPEG/ITU-T codec) and transmit video streams to a remote computing server for recognition or authentication purposes. A group of M devices in the same wireless local area network (WLAN) comprises a *wireless cluster*. A server running openstack or Hadoop (or a similar runtime environment suitable for cloud computing) [4] is used for analyzing visual data from numerous wireless clusters, as well as other computing tasks unrelated to object or face recognition. Each device can adapt the encoding bitrate, as well as the number of frames to produce (with the ensemble of N such settings comprising set $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$), in order to alleviate the impact of contention in the WLAN. At the same time, the visual analy-

The material is based upon work funded by the US Air Force Research Laboratory (AFRL). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not reflect the views of AFRL. OA, CT and MvdS acknowledge funding from NSF grant CNS-1016081. YA acknowledges funding from EPSRC grant EP/K033166/1.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

sis performed in the cloud can be adapted to scale the required processing time to alleviate the impact of task scheduling congestion in the cloud [5, 4], with the sets of contention and congestion levels represented by the discrete sets $\mathcal{T}$ and $\mathcal{G}$, respectively. In return, each device receives from the cloud a label that describes the recognized object or face (e.g. the object or person's name), or simply a message that the object or person could not be recognized. In addition, each device or wireless cluster can also receive feedback on the experienced WLAN medium access control (MAC) layer contention and the cloud task scheduling congestion conditions.

Thus, the "reward" for each device is the recognition result at each time step. Given that each wireless access point and the cloud computing infrastructure serves many more requests than the ones from a given cluster of devices (as illustrated in Figure 1), we can safely assume that for each device, the wireless contention and cloud congestion level are both independent of the actions taken by the devices within its cluster. This makes each device independent, since the decisions made by other devices do not affect the reward.

1.1. Relation to Prior Work

Each mobile device of Figure 1 seeks to maximize its own expected recognition rate at the minimum possible cost in terms of utilized wireless resources (i.e., MAC superframe transmission opportunities used). To this end, several approaches have been proposed that are based on reinforcement learning [6], such as Q-learning [5]. In these, the goal is to learn the state-value function, which provides a measure of the expected long-term performance (utility). However, they incur large memory overheads for storing the state-value function and they are slow to adapt to new or dynamically changing environments. A better approach is to intermittently explore and exploit when needed, in order to capture such changes. Index policies for multi-armed bandit (MAB) problems, contextual bandits [7][8], or epsilon-decreasing algorithms [9] can be used for this task. However, all existing bandit frameworks do not take into consideration the contention and congestion conditions as contexts in the application under consideration.

1.2. Contribution

Due to the lack of efficient methods that fully capture the problems related to online learning in multi-user wireless networks and cloud computing systems with uncertain and highly-varying resource provisioning, we propose a new online systematic learning theory based on multi-user contextual bandits, a natural extension of the basic MAB framework. We provide analytic estimates to compare its efficiency against the complete knowledge (or "oracle") benchmark in which the expected reward of every choice is known by the learner. Unlike Q-learning [6] and other learning-based methods, we prove that the regret bound—the loss incurred by the algorithm against the best possible decision that assumes full knowledge of contention and congestion conditions—is logarithmic if users do not collaborate and each would like to maximize their own utility. Finally, the proposed contextual bandit framework is general, and can also be used for learning in other wireless video applications that involve offloading of various processing tasks.

2. FORMALISM, ALGORITHM AND ANALYSIS

For each time instant, k, the mobile devices receive the contention and congestion levels in the wireless MAC and cloud scheduling, $t\left(k\right) \in \mathcal{T}$ and $g\left(k\right) \in \mathcal{G}$, respectively, and would like to find the best transmission setting to maximize their expected recognition rate. Under a standard algorithm for object or face recognition, such as the 2D PCA [10], the recognition rate varies based on: *(i)* the contention and congestion levels; *(ii)* the transmission settings $a\left(t\left(k\right), g\left(k\right)\right)$, $a\left(t\left(k\right), g\left(k\right)\right) \in \mathcal{A}$, selected by each device m, $m \in \mathcal{M}$. Once the compressed video is received and analyzed by the cloud, the device receives either the correctly-recognized object or person, or a response that the system is unable to recognize reliably based on the given content. In the latter case, the device repeats the recognition task until the object or person is recognized or the user cancels the task.

Let $\pi(t\left(k\right), g\left(k\right), a)$ be the expected recognition rate of the mth device with transmission settings a, given the contention and congestion levels $t\left(k\right)$ and $g\left(k\right)$ at the kth time instant, respectively[1]. The goal of each device is to explore the transmission settings in $\mathcal{A}$ and learn the expected recognition rate $\pi \in (0, 1)$ depending on the congestion level $g\left(k\right)$ and contention level $t(k)$. Note that it can then anticipate the average number of recognition attempts it will require in order to receive a recognition result with a predetermined confidence level. We will determine the performance of each learning algorithm in comparison to the optimal solution that selects the transmission setting $a^*\left(t\left(k\right), g\left(k\right)\right)$ yielding the highest expected recognition rate, given by

$$a^*\left(t\left(k\right), g\left(k\right)\right) := \arg\max_{\forall a \in \mathcal{A}} \pi\left(t\left(k\right), g\left(k\right), a\right). \quad (1)$$

The solution of (1) is defined as the oracle solution, since it assumes that all conditions for each case are precisely known beforehand. As a performance measure, we define the "regret" of a learning algorithm below.

Definition 1 (Regret). The regret after K iterations (time steps) is the loss incurred due to unknown system dynamics. For the mth device, the regret of the learning algorithm that selects the setting $a\left(t\left(k\right), g\left(k\right)\right)$ at each time instant k, $1 \leq k \leq K$, with respect to the best action is given by

[1] All the parameters defined in this paper are different for each mobile device $m \in \mathcal{M}$. For simplicity, we drop m subscript from our notation.

$$R(K) := \sum_{k=1}^{K} \pi\left(t\left(k\right), g\left(k\right), a^{*}\right) - E\left[\sum_{k=1}^{K} \hat{y}\left(a_{k}\right)\right] \quad (2)$$

with $\hat{y}(a) \in \{0, 1\}$ a discrete random variable modeling the recognition results received from the cloud under transmission setting a, and $E\left[\cdot\right]$ the statistical expectation. ■

2.1. Device-Oriented Contextual Learning

At any time step k, mobile device can be in one of the two following stages: *(i) exploration stage*, where it selects an arbitrary transmission setting to update the estimated recognition accuracy given the contention and the congestion levels; and *(ii) exploitation stage*, where mobile devices select the transmission setting yielding the highest estimated recognition accuracy given the WLAN contention level and the congestion level in the cloud. In this subsection, we focus on how learning is performed by one of the mobile devices, thus all the parameters defined below are for the specific mobile device m. However, all other mobile devices follow the same learning steps. Let $N_{t,g,a}(k)$ be the number of times transmission setting a is selected up to the kth time instant by the mobile device in response to the congestion and contention level $g\left(k\right)$ and $t(k)$ respectively. The mobile device checks if the following set is empty: $\mathcal{S}_{t(k),g(k)} = \{a \in \mathcal{A} : N_{t(k),g(k),a}(k) \leq c(k)\}$,where $c\left(k\right)$ is a deterministic control function that is monotonically increasing in k. In practice, $c\left(k\right)$ can be interpreted as the number of exploration steps required by the algorithm such that the deviation probability of the sample mean estimate of the expected reward of setting a decays with k^{-b} for some $b \geq 1$. The control function $c(k)$ controls if each transmission setting is explored sufficiently so that the sample mean of the recognition accuracies is accurate enough.

If $\mathcal{S}_{t,g} \neq \emptyset$, device m chooses an arbitrary transmission setting from this set and keeps the obtained recognition accuracy. If $\mathcal{S}_{t,g} = \emptyset$, this means that all the transmission settings are explored sufficiently. Then, each mobile device chooses transmission settings that yield the highest estimated recognition accuracy. Let $\mathcal{X}_{t,g,a}(k)$ be the set of recognition rates (a.k.a. set of rewards) obtained when selecting transmission setting a under WLAN contention level t and cloud congestion level g up to time k.. In addition, let $\alpha(k)$ be the optimized transmission setting at kth time instant according to:

$$\alpha(k) \in \arg\max_{a \in \mathcal{A}} \hat{Y}_{t,g,a}(k), \quad (3)$$

where $\hat{Y}_{t,g,a}(k)$ is the sample mean of the elements in $\mathcal{X}_{t,g,a}(k)$, i.e., $\forall r \in \mathcal{X}_{t,g,a}(k)$ with $0 < r < 1$ each recognition rate (or reward) obtained from the cloud:

$$\hat{Y}_{t,g,a}(k) = \sum_{r \in \mathcal{X}_{t,g,a}(k)} \frac{r}{|\mathcal{X}_{t,g,a}(k)|} \quad (4)$$

where $|\,\boldsymbol{.}\,|$ denotes the cardinality of a set. If there exists more than one setting maximizes (3), then the device simply selects one of them. The proposed algorithm for device-oriented contextual learning is given below.

```
Algorithm: Device-Oriented Contextual Learning
Input: $c(k)$; sets: $\mathcal{A}$, $\mathcal{G}$, $\mathcal{T}$
Initialization:
  $\hat{Y}_{t,g,a} = 0; \forall t \in \mathcal{T}\ , \forall g \in \mathcal{G}:\ N_{t,g,a} = 0; k = 1$
Repeat
  Get contention and congestion levels $t(k) \in \mathcal{T},\ g(k) \in \mathcal{G}$
  If $\exists a \in \mathcal{A}$ s.t. $N_{t,g,a}(k) \leq c(k)$
    Choose setting $a$
    Receive recognition rate (reward) $r_{t,g,a}$
    Update($N_{t,g,a}(k)$,$\hat{Y}_{t,g,a}$,$Y_{t,g,a}$)
  Else
    Find $\alpha(k) \in \arg\max_{a \in \mathcal{A}} \hat{Y}_{t,g,a}$
    Receive recognition rate (reward) $Y_{t,g,a}$
    Update($N_{t,g,\alpha(k)}(k)$,$\hat{Y}_{t,g,\alpha(k)}$,$Y_{t,g,\alpha(k)}$)
  End If
  $k \leftarrow k + 1$
End
Update($n$,$\hat{Y}$,$Y$):   $\hat{Y} \leftarrow \frac{n\hat{Y}+Y}{Y+1}$; $n \leftarrow n + 1$
```

Definition 2 (Suboptimality Gap and Minimum Suboptimality Gap). Let $\Delta_{t,g}(a^{-}) \triangleq \pi\left(t, g, a^{*}\right) - \pi\left(t, g, a^{-}\right)$ be the suboptimality gap of any transmission setting a^{-}, with $a^{-} \in \mathcal{A} \setminus a^{*}$, and its corresponding optimal setting $a^{*}(t, g)$ given by (1). We define the minimum suboptimality gap $\Delta_{\min}$ as the minimum difference between the expected recognition accuracy of the best transmission setting and second-best transmission setting, i.e., $\forall t \in \mathcal{T}$, $\forall g \in \mathcal{G}$, $\forall a^{-} \in \mathcal{A} \setminus a^{*}$: $\Delta_{\min} \triangleq \min \Delta_{t,g}(a^{-})$. ■

Lemma 1. If $\forall a^{-} \in \mathcal{A} \setminus a^{*}$, $\forall t \in \mathcal{T}$, $\forall g \in \mathcal{G}$: s.t.

$$\left|\hat{Y}_{t,g,a^{-}}(k) - \pi(t, g, a^{-})\right| \leq \frac{1}{2}\Delta_{\min}, \quad (5)$$

then: the optimized transmission setting given in (3) is $a^{*}\left(t, g\right)$ given in (1).

Proof: We have $|\hat{Y}_{t,g,a^{-}}(k) - \pi(t, g, a^{-})| < \frac{1}{2}\Delta_{\min}$; in the worst case, $\pi\left(t, g, a^{*}\right) - \hat{Y}_{k}\left(t, g, a^{*}\right) < \frac{1}{2}\Delta_{\min}$ and, for any suboptimal a^{-}, i.e., $\forall a^{-} \in \mathcal{A} \setminus a^{*}$: $\hat{Y}_{t,g,a^{-}}(k) - \pi\left(t, g, a^{-}\right) < \frac{1}{2}\Delta_{\min}$. Combining the last two inequalities with the fact that $\Delta_{t,g}(a^{-}) \leq \Delta_{\min}$ leads to: $\hat{Y}_{t,g,a^{*}}(k) - \hat{Y}_{t,g,a^{-}}(k) > 0$, which leads to the desired result. ■

Lemma 1 proves that, under accurate-enough estimates, the proposed algorithm will select the optimal transmission setting in the exploitations. We will use this to bound the suboptimal transmission setting selection in the exploitations.

2.2. Analysis of Device-Oriented Contextual Learning

The regret can be divided into two components. The first one is $R_e(K)$ the regret due to the explorations and $R_s(K)$ the regret due to suboptimal action selection in the exploitations. Since the rewards are bounded in $[0, 1]$, it is sufficient to bound the number of times that device chooses a suboptimal action. In the following lemmas, we will bound $R_e(K)$ and $R_s(K)$ separately.

Lemma 2. If $c(k) = 4\frac{b \ln k}{(\Delta_{\min})^2}$ for some $b > \frac{1}{2}$, then expected regret due to suboptimal action selection in exploitation step is: $E[R_s(K)] \leq 2N|\mathcal{G}||\mathcal{T}|H_K^{(2b)}$, with $H_K^{(2b)}$ the Generalized Harmonic Number.

Proof: The proof is based on using the Chernoff-Hoeffding inequality to derive (details omitted due to space limitations)

$$E[R_s(K)] \leq \sum_{k=1}^{K} \sum_{\forall t,g,a} P\left(\left|\hat{Y}_{t,g,a}(k) - \pi\left(t,g,a\right)\right| \right.$$
$$\left. \leq \frac{1}{2}\Delta_{\min}, N_{t,g,a}\left(k\right) \geq c\left(k\right)\right)$$

where $P(e)$ denotes the probability of event e. The last expression is upper bounded by $2N|\mathcal{G}||\mathcal{T}|\sum_{k=1}^{K} k^{-2b}$, with the summation term being $H_K^{(2b)}$ [11]. ■

Lemma 3. Under the conditions of Lemma 2, the regret due to explorations is $E[R_e(K)] \leq |\mathcal{G}||\mathcal{T}|N\left(1 + 4\frac{b \ln K}{(\Delta_{\min})^2}\right)$.

Proof. At any time k, at most $c(k) + 1$ exploration steps took place for each $g \in \mathcal{G}$. This leads to: $E[R_e(K)] \leq \sum_{\forall t,g,a}\left(1 + 4\frac{b \ln K}{(\Delta_{\min})^2}\right) = |\mathcal{G}||\mathcal{T}|N\left(1 + 4\frac{b \ln K}{(\Delta_{\min})^2}\right)$. ■

Theorem 1. *Under the conditions of Lemma 2 and $b = 1$,*

$$E\left[R\left(K\right)\right] \leq |\mathcal{G}|\left|\mathcal{T}\right| N\left(1 + 4\frac{b \ln K}{\Delta_{\min}^2} + 2H_K^{(2)}\right) \quad (6)$$

Proof. We have: $E\left[R\left(K\right)\right] = E\left[R_e\left(K\right)\right] + E\left[R_s\left(K\right)\right]$, which, from Lemmas 2 and 3 and with $b = 1$, is upper bounded by the desired result. ■

We proved that this algorithm can achieve the logarithmic regret, which is the lowest possible regret that can be achieved by any function $c(k)$ [12]. Moreover, this regret implies that $\lim_{K\to\infty} \frac{R(K)}{K} = 0$, i.e., the time-averaged regret leads to zero and the total expected recognition rate will thus converge to the value of the oracle solution. The logarithmic regret bounds can still be achieved for dependent devices scenario when the cloud learns the setting profiles for the devices and reccomends the transmission settings for them.

3. NUMERICAL RESULTS

Our simulation environment comprises 4 mobile devices connected via a IEEE 802.11 WLAN to a cloud-computing server. Videos of human faces are produced by random images of persons taken from the extended Yale Face Database B (39 cropped faces of human subjects under varying illumination). Each video comprises 34 images from the same person and it is compressed to a wide range of bitrates via the H.264/AVC codec (x264 codec, crf $\in \{4, 14, 24, 34, 44, 51\}$). The 2D PCA algorithm [10] is used at the cloud side for face recognition from each decoded video (with the required training done offline as per the 2D PCA setup [10]). More than 80% of the video frames have to match to the same person in the database to declare this video as "recognized". There is a time window set for recognition, which limits the number of frames received by the cloud under varying WLAN contention levels (delay is increased under contention due to the backoff and retransmissions of IEEE 802.11 WLANs). Similarly, because of randomly varying congestion in the cloud, only a limited number of the received video frames is actually used by 2D PCA, thereby affecting the recognition rate.

Table 1 presents the average number of retries performed per recognition action by our method (with and without using the cloud congestion information as context) in order to achieve recognition rate of 90%. We also present results obtained by: *(i)* the optimal setting of (1) that assumes full system knowledge (oracle bound); *(ii)* Q-learning [6, 13]. The results indicate that, after 250 recognition attempts (each attempt comprises the retries listed), our algorithm approaches the oracle bound and, for the same recognition rate, incurs less retries per attempt in comparison to Q-learning.

4. CONCLUSIONS

We propose a contextual bandit framework for learning contention and congestion conditions in object or face recognition via wireless mobile streaming and cloud-based processing. Analytic results show that our framework converges to the value of the oracle solution (i.e., the solution that assumes full knowledge of congestion and contention conditions). Simulations within a cloud-based face recognition system demonstrate it outperforms Q-learning, as it quickly adjusts to contention and congestion conditions.

Table 1. Average attempts (with the oracle bound given in parentheses) to obtain recognition rate 0.9 with 2D-PCA.

Iteration Method	$T = 50$	$T = 100$	$T = 250$	$T = 1000$
Proposed	3.3 (1.7)	3.1 (1.6)	2.4 (1.5)	1.9 (1.5)
Proposed no context	3.1 (1.7)	2.8 (1.6)	2.6 (1.6)	2.4 (1.6)
Q-learning	3.5 (1.7)	2.8 (1.6)	2.7 (1.5)	2.2 (1.5)

5. REFERENCES

[1] D. Siewiorek, "Generation smartphone," *IEEE Spectrum*, vol. 49, no. 9, pp. 54–58, 2012.

[2] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, "Mobile visual search," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011.

[3] B. C. Becker and E. G Ortiz, "Evaluation of face recognition techniques for application to facebook," in *8th IEEE Internat. Conf. on Automatic Face & Gesture Recognition, 2008. FG'08.* IEEE, 2008, pp. 1–6.

[4] A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in *Proc. 10th ACM SIGCOMM Conf. on Internet Meas.* ACM, 2010, pp. 1–14.

[5] Shaolei R. and M. van der Schaar, "Efficient resource provisioning and rate selection for stream mining in a community cloud," *IEEE Trans. on Multimedia*, vol. 15, no. 4, pp. 723–734, 2013.

[6] Barto A. Sutton, R., Ed., *Reinforcement learning, an introduction.*, Cambridge: MIT Press/Bradford Books, 1998.

[7] A. Slivkins, "Contextual bandits with similarity information," in *in 24th Annual Conference on Learning Theory (COLT)*, 2011.

[8] John Langford and Tong Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information.," in *Advances in neural information processing systems*, 2007.

[9] Nicolo Cesa-Bianchi Auer, Peter and Paul Fischer, "Finite-time analysis of the multiarmed bandit problem.," *Machine learning*, vol. 47, pp. 235–256, 2002.

[10] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[11] Knuth-D. E. Graham, R. L. and O. Patashnik, Eds., *Concrete Mathematics: A Foundation for Computer Science, 2nd ed.*, Reading, MA: Addison-Wesley, 1994.

[12] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv Appl Math*, vol. 6, 1985.

[13] X. Zhu, C. Lany, and M. van der Schaar, "Low-complexity reinforcement learning for delay-sensitive compression in networked video stream mining," in *IEEE Internat. Conf. on Multimedia and Expo (ICME), 2013*. IEEE, 2013, pp. 1–6.