Towards Resource-efficient Wireless Edge Analytics for Mobile Augmented Reality Applications

Livia Elena Chatzieleftheriou[1], Georgios Iosifidis[2], Iordanis Koutsopoulos[1], Douglas Leith[2]
[1]Athens University of Economics and Business, [2]Trinity College Dublin

***Abstract*—From entertainment to education, augmented reality (AR) is about to impact positively our everyday lives. Enhanced capabilities of mobile devices, such as smartphones or wearables, as well as ubiquitous network connectivity give AR the opportunity to prosper. Despite these improvements, AR requires computationally heavy tasks, such as context recognition and classification through image or video processing, which are hard to fulfill on mobile devices. To this end, solutions for computation offloading to cloud servers have been proposed. We consider a scenario where context identification is performed through elicitation of user-generated information, such as images or small video files. It is the quantity of this information that ultimately determines the context classification precision, which we model as a Binomial random variable. We introduce the problem of maximizing a lower bound of the precision of context classification through prudent resource allocation, namely computation offloading, and bandwidth and computational capacity allocation at the wireless network edge. We define the context classification precision as a function of the quantity of information that users provide, and we demonstrate through numerical experiments that appropriate management of the limited resources at the wireless edge can maximize the classification precision of data analytics mechanisms needed for augmented reality applications.**

***Index Terms*—Mobile Augmented Reality, Mobile Computing, Computation Offloading, Edge Analytics, Classification Precision, Context Awareness**

I. Introduction

Augmented reality (AR) is rapidly approaching our everyday life, and it is increasingly used in various applications and fields: from tourism and navigation to entertainment and advertisement, up to training and education. AR is interactive in real time, aligning real and virtual objects with each other. A special case of AR is Mobile Augmented Reality (MAR), which is realized through the deployment of a mobile app [1].

In order to reap the benefits of AR applications and have good Quality of Experience (QoE) for the user, *i.e.*, to project the right virtual object onto a physical object, a central issue is that of context identification and classification. In [2], the authors examine the relationship between tourist satisfaction and the perceived quality of AR applications, concluding that users are more concerned with high-quality content and a good degree of personalized service than system quality. Next generation MAR systems will incorporate recommender systems which will use contextual information from the surroundings of the user.[1] Determining the context of users is a complex task that involves object recognition on mobile devices, by using information both from sensors on the devices and information explicitly given from users, such as images or small videos. Context identification is a major machine-learning objective outcome which needs to be performed in real-time and under limited edge resources, in order to timely project the right virtual object to users.

However, precision in context classification is costly and challenging to guarantee, given the limited computational power and bandwidth resources at the edge. This naturally raises challenges, as computationally expensive tasks have to be performed in real-time. Since mobile devices have limited computational resources, offloading of computationally-intense tasks to a cloud server may be needed occasionally. Furthermore, bandwidth is also scarce. Hence, in order to fill the gap between the insufficient mobile processing capability and the high computation demands of AR applications, *there is a need for a sophisticated mechanism which will take decisions regarding the amount of context information that needs to be collected from the users and the location (device or cloud) to process this information.*

In this work, we consider a single cell with many users, each user running an AR mobile app. For example, consider an app showing short videos containing information about archaeological findings the user is pointing at. In order to project appropriate virtual objects on top of the real objects to which the users are pointing, not only location but also information such as the noise levels or the luminosity of the environment of the user is needed, *i.e.*, the users' context needs to be identified and classified. Context identification relies on a variety of voluminous and computation-intense information items, such as videos or images that are solicited by the user. Classification precision is determined by the quantity of the information submitted by the users. Further, video analytics is a particularly resource-demanding task which needs to be performed in real-time. Mobile devices have limited computational power, and the shared uplink bandwidth and energy resources at the edge are also limited. Hence, the question arises *what volume of information items to request from users and where to do the processing of information items so as to*

Corresponding author: Livia Elena Chatzieleftheriou (liviachatzi@aueb.gr)
This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 17/CDA/4760. The research work of L.E. Chatzieleftheriou and I. Koutsopoulos was supported by the European Commission H2020 Research Program under Grant Number 688768 - netCommons (Network Infrastructure as Commons).

[1]Context-aware recommender systems are usually outperforming traditional (context-unaware) RS regarding prediction accuracy of users' preferences, resulting in better quality recommendations [3].

have accurate data analytics. This is the question we attempt to answer in this paper.

A. Our contribution

The contributions of our paper are as follows.

- We develop a model for user context and how this could be obtained through costly data collection from users.
- We model the classification precision as a function of the volume of information obtained by users so as to capture the realistic limitations in obtaining sufficient information from users in order to infer their context.
- We formulate a mathematical optimization problem for maximizing a lower bound on the precision of context classification which we derive analytically, under constraints related to energy in devices and wireless bandwidth, as well as stringent delay constraints of AR applications. We aim in jointly deciding the total volume of requested information and the location (remote or local) where the process will take place.
- We perform simulation studies to investigate our model's underlying tendencies. We conclude that taking jointly decisions regarding *where* to process information and *how much* quantity of information to process, leads to optimum resource allocation. Indicatively, a joint consideration can lead to a 35% higher lower bound for classification precision, compared to simple algorithms considering separately these parameters. Moreover, we evidence that a less stringent delay constraint is more important for classification precision than lower energy consumption.

The rest of the paper is organized as follows. In section II we describe the model, and in sections III and IV we state the problem and investigate on its properties, and then we evaluate our model respectively. In section V we refer to related work, and in section VI we present our conclusions and future work.

II. System Model

A. Model Components

We depict our model in Fig. 1. We consider a single base station (BS) with unlimited computational and energy resources. We consider that the wireless link between the BS and the users has limited bandwidth R^B for receiving and for transmitting data from/to AR users' devices.

We assume a set of users $u \in \mathcal{U}$ who are in range with the BS, each equipped with an AR device. Each AR device associated with user u is equipped with sensors that give information to the system, *e.g.*, geographic location and battery levels of the device in use. The application provider decides in advance the maximum amount R_u^E of energy resources that can be used by the application, so that the user is not discouraged by the energy consumption of the application and remains engaged. Devices also have limited computational capacity.

We next define the users' context. Context types refer to the company, mood, or other attributes that characterize the user's spatiotemporal state [3]. For example, consider the set

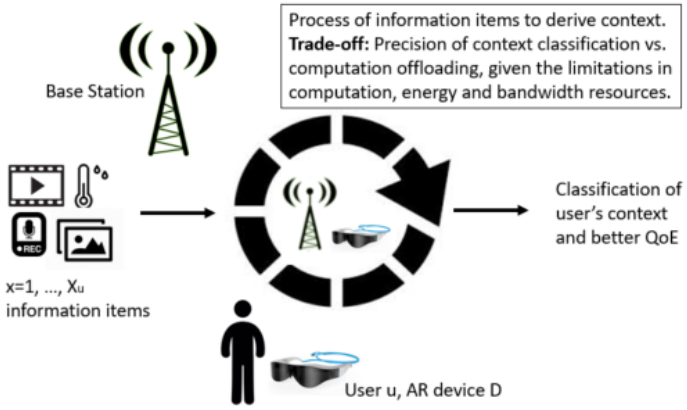


Fig. 1. Illustration of the AR model and basic components. Information items are solicited from users in order to better classify their context. Context classification process is a resource-demanding task and, to overcome this, computation offloading to the BS is proposed. Decisions are taken about the volume of information items to solicit from user, and the portion of the computational task to process locally at device or offload to to be processed at the BS.

$\mathcal{N}$ of context types, $\mathcal{N} = \{\text{Mood, Company, Location, Objects of focus}\}$. Each element in this set is a factor that influences the suitability of projection of an AR item to the users and, ultimately, her Quality of Experience (QoE) [3]. For simplicity, we consider only one context category, so $|\mathcal{N}| = 1$, and we let $\mathcal{L} = \{1, \ldots, L\}$ be the set of discrete context labels that matter for the application. For example, assume that we are interested in classifying the scenery where the user is pointing at, *i.e.*, the real object over which the virtual object will be projected, and let $L = 2$, and $\mathcal{L} = \{\text{building, statue}\}$.

B. Context and Information Items

1) Context Information Items: In order to better identify the users' context and provide them with a higher QoE, our system requires users to submit some "information items" that will be used to classify their context. The system gets the requested information items, either implicitly (*e.g.*, from sensors) or explicitly (*e.g.*, requesting it from users) [1]. For simplicity, in this paper we assume that the users' context is defined by one type of information items. For example, assume that in order to classify the object of focus, our system requests only short videos [4]. Let $x_u \in \{1, 2, \ldots, X_u\}$ be the variable describing the volume of information items user u submits, *e.g.*, the number of frames, and X_u be the maximum volume she is willing to share, *e.g.*, for privacy reasons [5].

2) Processing costs of information items: Due to different characteristics of different devices, we consider C_u^E and $C_{u,D}^P$ to be the device-specific energy and computational power costs per unit of volume of information, *e.g.*, per frame, for obtaining and processing information items of user u in her device (*e.g.*, as in [5]), and $C_{u,B}^P$ the computational power cost per unit of volume of information at the server. We consider the computational cost in seconds [4]. In terms of bandwidth, the cost of processing an item at the BS consists in the cost of transmitting it, *e.g.*, transmitting fingerprints or entire frames

(*e.g.*, as in [6]). We denote as C_u^B the per unit of volume, *e.g.*, per frame, cost for sending information items of user u to the BS, *e.g.*, in bits.

C. Context classification and precision

1) Classification: Taking as input volumes x_u, the context label $l_u \in \mathcal{L}$ for each user u is inferred. Given an appropriately trained classifier, *e.g.*, an SVM, the output is a class membership, *i.e.*, one specific label in $\mathcal{L}$. In the previous example, the output will be either "building" or "statue". The real-time requirements of MAR applications suggest that the task has to be completed within a maximum delay T^D.

2) Precision of Classification: We denote as p the precision of our classifier, defined as the percentage of the true positives divided by the percentage of both true and false positives. To capture the heterogeneity in how users submit information items, our model considers classification precision to be user-dependent. Moreover, recall that different computation costs, $C_{u,D}^P$ and $C_{u,B}^P$, occur for classifications performed locally and remotely. Due to the limited computational resources of mobiles devices and the stringent delay requirements for object classification on mobile devices [7], [4], we assume that different classification algorithms (with the same delay cost C_u^D per unit of volume of information items) run at mobile devices and at the BS. This is justified by existing works, *e.g.*, [5], [4], and results in different classification precision for on-device (local) or cloud (remote) classification. The classification precision of the already trained classifier in use for a single information item is a random variable P_{ur} with mean $p_{ur} \in [0,1]$, where $u \in \mathcal{U}$ and $r \in \{B, D\}$ the location where computation is executed, where $r = B$ for remote and $r = D$ for on-device classification.

We consider the classification precision to depend also on the volume of information items provided by the user: The correct classification of each submitted information item, *e.g.*, frame, is considered a Bernoulli trial, with probability of success p_{ur}. Naturally, the total number W_{ur} of correct classifications when x_u information items are processed follows a Binomial distribution $B(x_u, p_{ur})$. The final label $l_u \in \mathcal{L}$ for user's u context is defined by the majority of multiple Bernoulli trials, *i.e.*, $\mathbf{P}[l_u \text{ is correct}] = \mathbf{P}[W_{ur} > \frac{x_u}{2}]$.

Let $Z_{ur} = x_u - W_{ur}$ be the number of wrong classifications in user's u context when x_u information items are available. Clearly, $Z_{ur} \sim B(x_u, 1-p_{ur})$, with $\mathbf{E}[Z_{ur}] = x_u(1-p_{ur})$ and $\mathbf{Var}[Z_{ur}] = x_u p_{ur}(1-p_{ur})$. Then, $\mathbf{P}[W_{ur} > \frac{x_u}{2}] = \mathbf{P}[x_u - Z_{ur} > \frac{x_u}{2}] = \mathbf{P}[Z_{ur} < \frac{x_u}{2}] = 1 - \mathbf{P}[Z_{ur} \geq \frac{x_u}{2}] = 1 - \mathbf{P}[Z_{ur} \geq x_u(1-p_{ur}) + x_u(p_{ur} - 1/2)] = 1 - \frac{1}{2}\mathbf{P}[|Z_{ur} - \mathbf{E}[Z_{ur}]| \geq x_u(p_{ur} - 1/2)]$. Using Chebychev's inequality, it is $1 - \frac{1}{2}\mathbf{P}[|Z_{ur} - \mathbf{E}[Z_{ur}]|] \geq 1 - \frac{1}{2}\frac{\mathbf{Var}[Z_{ur}]}{\left(x_u(p_{ur}-1/2])\right)^2}$ $= 1 - \frac{1}{x_u}\frac{p_{ur}(1-p_{ur})}{2(p_{ur}-1/2)^2}$.

Finally, it is $\mathbf{P}[l_u \text{ is correct}] \geq g_u^r(x_u)$, where

$$g_u^r(x_u) = 1 - \frac{1}{x_u}\frac{p_{ur}(1-p_{ur})}{2(p_{ur} - 1/2)^2} \tag{1}$$

is a lower bound for the classification precision of user's u context classification in location $r \in \{B, D\}$ when $x_u > 0$ information items, *e.g.*, frames, are processed. In real scenarios, values p_{ur} can be found from historical data collected by the application provider.

3) Properties of Precision Certainty: From (1) we can trivially conclude that as the volume x_u of information items submitted by user u increases, (i) the precision increases and (ii) the rate of increase decreases, *i.e.*, precision is a monotone increasing and concave function of the volume x_u of submitted information items. This follows intuition as, for example, two seconds of video are likely to give more information about the context than one-second video, *e.g.*, more details about the scenery the user is pointing at. This results in higher precision but with diminishing returns, as from the first second of video we already have some knowledge about the user's context.

III. Computation offloading for precision maximization

A. Problem Statement

In order to provide users with high quality AR recommendations, information items need to be processed and their context needs to be classified. There is clearly a trade-off between data usage, energy consumption, delay and classification precision uncertainty: classification at the devices reduces data usage but increases uncertainty and energy consumption.

We assume that the process of an information item can be split to partially run on the device and the BS at the same time. We denote as $\mathbf{x^B}$ and $\mathbf{x^D}$ the vectors whose u-th elements are the decision variables x_u^B and x_u^D, interpreted as the volume of information items that is requested from user u and processed at the BS or locally, respectively. Clearly, $x_u = x_u^B + x_u^D \leq X_u$ is the total quantity of information items requested by user u. We consider that the final precision regarding user u is the average of the precisions obtained by classifications done locally and remotely[2] and consider the continuous analogous of the defined precision functions in (1).[3] We define

$$g(\mathbf{x^D}, \mathbf{x^B}) = \sum_{u \in \mathcal{U}} g_u^B(x_u^B) + \sum_{u \in \mathcal{U}} g_u^D(x_u^D) \tag{2}$$

and consider the following optimization problem:

$$\max_{\mathbf{x^D}, \mathbf{x^B}} g(\mathbf{x^D}, \mathbf{x^B}) \tag{3}$$

s.t.

$$x_u^D \leq \min\left\{\frac{R_u^E}{C_u^E}, \frac{T^D}{C_u^D}\right\}, \quad \forall u \in \mathcal{U} \tag{4}$$

$$\sum_{u \in \mathcal{U}} x_u^B C_u^B \leq R^B, \tag{5}$$

$$x_u^B \leq \frac{T^D}{C_u^D}, \quad \forall u \in \mathcal{U} \tag{6}$$

$$x_u^D + x_u^B \leq X_u, \quad \forall u \in \mathcal{U} \tag{7}$$

[2]Since the computation task is split, we can make application-specific assumptions regarding the final precision of classification. For example, we can assume that the combination of the two classifiers results in the maximum of the two obtained precisions, or a weighted sum of them.

[3]Continuous extensions of $g_u^r(\cdot)$ can be inferred, *e.g.*, when variables x_u denote the seconds of video, with interpolation or extrapolation methods.

$$0 < x_u^D, x_u^B \leq 1, \quad \forall u \in \mathcal{U}. \tag{8}$$

This problem consists in maximizing a *lower bound* on the probability of correct classification of the users' context with respect to the volume and the location (remote or local) of process of the users' information items, and not the probability of correct classification itself, which would constitute a much more complex (computationally) metric to derive.

Equation (4) captures energy and delay limitations for classifications executed on mobile devices. Equations (5) and (6) capture network (*e.g.*, bandwidth) and delay limitations for classification done remotely. Equations (7) ensure that the requested volume of information items will not exceed X_u and (8) set upper and lower bounds to x_u^B and x_u^D.

B. Properties of the Problem

Function $g(\mathbf{x^D}, \mathbf{x^B})$ is monotone increasing and concave in the volumes x_u^D and x_u^B of request of information items, $\forall u \in \mathcal{U}$. This trivially holds, as our objective is the sum of monotone increasing and concave functions (section II-C3).

This problem consists in the maximization of a non-linear convex monotonically increasing function under linear constraints. Since the domain $(0, X_u]^{2\,|\mathcal{U}|}$ of $g(\mathbf{x^D}, \mathbf{x^B})$ is a convex and bounded set and $g(\mathbf{x^D}, \mathbf{x^B})$ is monotonically increasing, every local maximum is a global maximum. Given the properties of the problem, the optimal solution can be found numerically by applying the KKT conditions [8].

IV. Evaluation

A. Synthetic dataset

We consider a system with 100 users and experiment with several values of energy consumption and delay tolerance. We experiment with the lower bound we presented in (1). We consider higher precision for classifications executed at the BS and lower precision for those executed on mobile devices. Following the results in [5], we assume a ratio of $2/3$ for expected precision is obtained at the BS and the devices. Following results from [7], we consider high and low delay tolerance T^D to allow the 75% and 10% of the maximum volume of information items, respectively, to be processed within real-time constraints of MAR applications. Similarly, we consider high energy consumption for classifications done locally to allow only 10% of the maximum volume of information items to be processed locally, as opposed to the 75% we assume for low energy consumption. All results consider the percentage of the bandwidth that would be needed to offload the maximum volume of information items for all users.

B. Schemes under Comparison

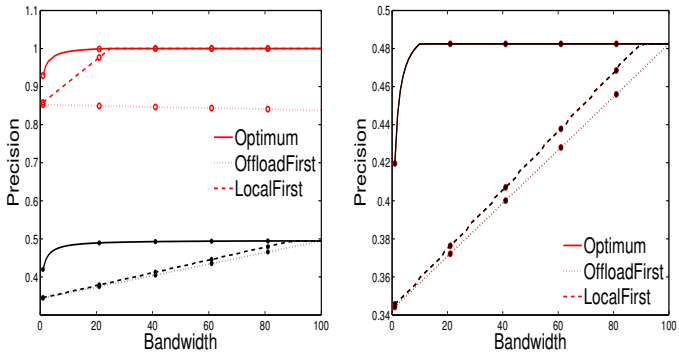
In real-life scenarios, many application providers may use simplistic schemes to address the computation offloading of context classification, thus being mindful about extra delay costs. The convex optimization we propose can be solved easily by existing efficient solvers. In our evaluation we consider two simple online heuristics that do not present any further delay cost to find a solution. Namely:

OffloadFirst: Given bandwidth and delay constraints assumes a fixed order of users and offloads the maximum possible intensity for each user, until constraints are tight. Then process the remaining quantities locally.

LocalFirst: Given energy and delay constraints processes locally the maximum possible quantity of information items for each user. The remaining volume is offloaded to the BS.

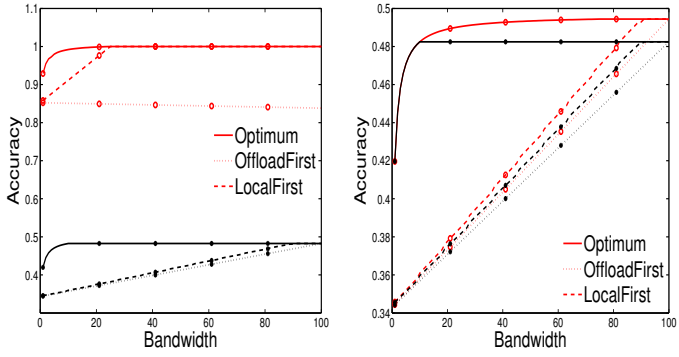
Both schemes aim at maximization of the lower bound of correct classification. The quantity of information items that is processed locally is limited due to delay and energy constraints while the processed at the BS quantity is limited due to delay and bandwidth constraints in all schemes.

C. Numerical Results and Derived Conclusions



(a) High delay tolerance (b) Low delay tolerance

Fig. 2. Classification precision lower bound *vs.* network resources. A lower bound for classification precision is depicted for low 'o' and high '*' energy consumption for context classification on the device, for (a) high and (b) low delay tolerance.



(a) Low energy consumption (b) High energy consumption

Fig. 3. Lower bound of classification precision *vs.* network resources. Lower bound for classification precision is depicted for high 'o' and low '*' delay tolerances for (a) low and (b) high energy consumption on mobile devices due to classification.

1) Relation of schemes and their limitations: Taking into consideration the definitions of the heuristics and from Figs. 2 and 3, we conclude that:

(i) Due to the fixed order considered on users by OffloadFist, its maximum lower bound is always lower than the proposed, both in terms of classification precision and in terms of network resources requirement, even if enough bandwidth resources are available to transmit all information items for all users.

(ii) Due to the concave objective function, LocalFirst's maximum lower bound is always lower than the proposed, both in terms of classification precision and in terms of network resources requirement, even if enough energy resources are available to process all volume of information items locally.

(iii) As opposed to LocalFirst and OffloadFirst, our scheme reaches the maximum possible value, *i.e.*, $\mathbf{P}[l_u \text{ is correct}] = 1$, even for very restricted, both energy and bandwidth, available resources.

2) About the influence of constraints: Lower delay tolerance results in lower classification precision (Fig. 3) and higher per volume energy consumption results in lower classification precision (Fig. 2). Also, comparison of the obtained precision in Figs. 2b and 3b indicates that a delay-related limitation is more restrictive than an energy-related limitation.

Finally, taking jointly decisions leads to optimum resource allocation and outperforms both OffloadFirst and LocalFirst. The gain over the comparison schemes significantly depends on the restrictions about delay tolerance and energy consumption. All precision values in Figs. 2 and 3 are normalized with respect to the maximum value of the lower bound for classification when delay tolerance and energy constraints are idle.

V. Related Work

In [2], the authors conclude that content quality and personalized service quality have a stronger effect on users' QoE than system quality. Along these lines, tour guide solutions for mobile users which incorporate in their decisions contextual information have been developed, *e.g.*, [9]. In [10], the authors recommend items based on the focus of the user and her distance from them. Nonetheless, context recognition requires computationally expensive tasks. According to [4], it takes on average more than 2 seconds to finish object recognition on a mobile CPU, which is an intractable delay for a real-time AR application. In [7], the authors measure cloud offloading for visual tasks (recognition) for AR, concluding that cloud offloading is a promising technology to fill the gap between the insufficient mobile processing capability and the high computation demands of AR applications. Towards a less bandwidth-consuming object recognition, the authors in [6] take fingerprints of images that users get to offload, in order to process the recognition task in the cloud. In [11], the authors manage cloud resources for offloading requests to both improve offloading performance seen by mobile devices and reduce the monetary cost per request to the provider. In [12], the authors implemented a MAR system based on cloud computing. This system uses a smartphone to capture images and sends processed features to the cloud. The mobile device is used to perform some image processing tasks and the cloud is used to realize heavy computations.

VI. Conclusions and future work

We aim at resource-efficient edge analytics for MAR applications and develop a mathematical model for context classification, taking into consideration the realistic limitations in obtaining sufficient data from users. After taking into account energy, bandwidth and delay constraints of real-time MAR devices and applications, we formulate a mathematical optimization problem for maximizing context classification precision, we analyze its properties and characterize its solution.

We conclude that taking jointly decisions regarding the location (remote or local) of the classification and the quantity of data to use for this, leads to maximization of the lower bound of the obtained classification precision and optimum resource allocation at the wireless edge. Moreover, there is evidence that a less stringent delay constraint is more important than lower energy consumption.

In our future work, we plan to enrich our model, considering that context is described by more than one category, *e.g.*, considering both the classification of visual and modal context of users. Moreover, we will investigate the influence of more than one information items on each context category, *e.g.*, both images and sound recordings that can be used to determine the users' context.

References

[1] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile augmented reality survey: From where we are to where we go," *IEEE Access*, 2017.

[2] T. Jung, C. NamHo, and M. C. Leue, "The determinants of recommendations to use augmented reality technologies: the case of a korean theme park.," *Tourism Management*, vol. 49, pp. 75–86, 2015.

[3] G. Adomavicius, M. B., R. F., and T. A., "Context-aware recommender systems," in *AI magazine*, vol. 32(3), pp. 67–80, 2011.

[4] T. Y. Chen, L. Ravindranath, S. Deng, P. Bahl, and H. Balakrishnan, "Glimpse: Continuous, real-time object recognition on mobile devices," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys 2015, Seoul, South Korea, November 1-4, 2015*, pp. 155–168, 2015.

[5] S. Fan, T. Salonidis, and B. C. Lee, "A framework for collaborative sensing and processing of mobile data streams: demo," in *MobiCom*, pp. 501–502, ACM, 2016.

[6] P. Jain, J. Manweiler, and R. Roy Choudhury, "Low bandwidth offload for mobile ar," in *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '16, (New York, NY, USA), pp. 237–251, ACM, 2016.

[7] W. Zhang, B. Han, and P. Hui, "On the networking challenges of mobile augmented reality," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, VR/AR Network '17, (New York, NY, USA), pp. 24–29, ACM, 2017.

[8] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2016.

[9] D. Gavalas, C. Charalampos Konstantopoulos, K. Mastakas, and G. Pantziou, "Mobile recommender systems in tourism," in *Journal of Network and Computer Applications*, vol. 39, pp. 319–333, 2014.

[10] D. Chatzopoulos and P. Hui, "Readme: A real-time recommendation system for mobile augmented reality ecosystems," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pp. 312–316, 2016.

[11] C. Shi, K. Habak, P. Pandurangan, M. Ammar, M. Naik, and E. Zegura, "Cosmos: Computation offloading as a service for mobile devices," in *Proceedings of the 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '14, (New York, NY, USA), pp. 287–296, ACM, 2014.

[12] B. R. Huang, C. H. Lin, and C. H. Lee, "Mobile augmented reality based on cloud computing," in *Anti-counterfeiting, Security, and Identification*, pp. 1–5, Aug 2012.