Reducing Response Time for Multimedia Event Processing using Domain Adaptation

Asra Aslam
Insight Centre for Data Analytics
NUI Galway, Ireland
asra.aslam@insight-centre.org

Edward Curry
Insight Centre for Data Analytics
NUI Galway, Ireland
edward.curry@insight-centre.org

ABSTRACT

The Internet of Multimedia Things (IoMT) is an emerging concept due to the large amount of multimedia data produced by sensing devices. Existing event-based systems mainly focus on scalar data, and multimedia event-based solutions are domain-specific. Multiple applications may require handling of numerous known/unknown concepts which may belong to the same/different domains with an unbounded vocabulary. Although deep neural network-based techniques are effective for image recognition, the limitation of having to train classifiers for unseen concepts will lead to an increase in the overall *response-time* for users. Since it is not practical to have all trained classifiers available, it is necessary to address the problem of training of classifiers on demand for unbounded vocabulary. By exploiting transfer learning based techniques, evaluations showed that the proposed framework can answer within ∼0.01 min to ∼30 min of response-time with accuracy ranges from 95.14% to 98.53%, even when all subscriptions are new/unknown.

CCS CONCEPTS

• **Information systems** → *Multimedia streaming*; • **Computing methodologies** → *Neural networks*; • **Software and its engineering** → *Publish-subscribe / event-based architectures.*

KEYWORDS

Domain Adaptation, Online Training, Internet of Multimedia Things, Event-Based Systems, Multimedia Stream Processing, Object Detection, Transfer Learning, Smart Cities, Machine Learning

ACM Reference Format:
Asra Aslam and Edward Curry. 2020. Reducing Response Time for Multimedia Event Processing using Domain Adaptation. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390722

1 INTRODUCTION

Due to ever increasing shift of data towards multimedia, the inclusion of "multimedia things" in the domain of Internet of Things (IoT) is a crucial step for the emerging applications of smart cities [2, 3, 24, 34, 42]. Event processing systems [12, 14] are designed to process data streams (consisting of mostly scalar data excluding multimedia data). In case of smart cities, multiple types of multimedia applications may require handling of multiple subscriptions belonging to multiple domains (like {car, bus, pedestrian, bike} ϵ traffic management, {car, taxi, bike} ϵ parking management, {ball, person} ϵ sports event management etc.). High performance requirement of real-time systems can be accomplished using existing image processing systems but they are designed only for specific domains, have limited user expressibility, and cannot successfully realize the goal of generalizable multimedia event processing due to their bounded object detection capability.

In our previous work [5, 6] we analyzed the problem of generalized multimedia event processing but recognized the requirement of availability of trained classifiers for unknown concepts/objects within subscriptions (unbounded vocabulary [52]). The online training of classifier on request of any new/unknown subscription is an option to be explored, which will help either in switching (transforming) from one classifier to another (like bus → car) or in the construction of completely new classifier (like ball). Also, existing DNN based techniques [18] are well-known for easy knowledge transfers among domains [16, 30, 46] but focused either on improving accuracy or testing time. They do not analyze the overall response time of the process of transfer and its impact on accuracy.

In this work, we propose an adaptive multimedia event processing model, that leverages transfer learning-based techniques for domain adaptation to handle unknown/new subscription within an acceptable time frame. An example of multimedia event processing specifically for the detection of objects is shown in Fig. 1. Main contributions of this article include a definition of quality metric "response-time" supporting "unknown subscriptions", an adaptive approach using online classifier construction to support multiple domain-based subscriptions, and an instantiation of a classifier learning model by transferring knowledge among classifiers using fine-tuning and freezing layers of neural network-based object detection models.

2 BACKGROUND WITH RELATED WORK

Very few multimedia event based architectures for Internet of Multimedia Things (IoMT) proposed in recent works [2, 42, 44], focused on scalability and multimodal big data. However augmenting IoT systems with multimedia event based approaches is not straightforward, and still haven't been combined yet as an end-to-end model.

2.1 Domain-Specific Event Recognition

Event recognition in multimedia is one of the popular areas of research [21, 32, 51]. Traffic recognition systems [17, 26] are highly

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390722

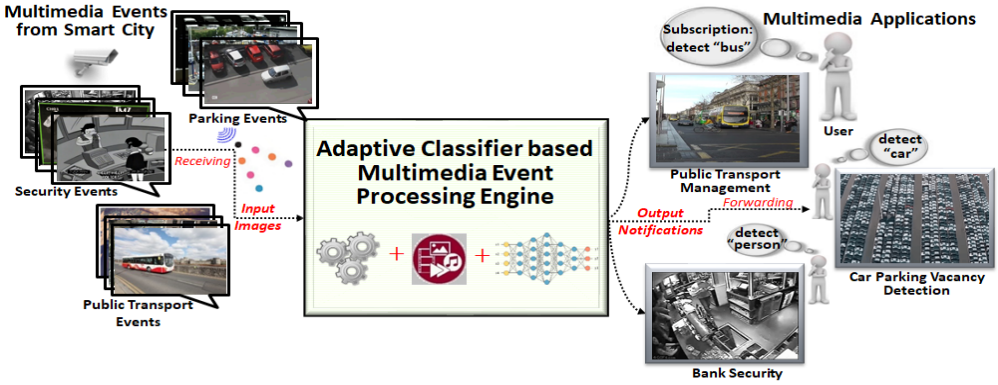


Figure 1: Generalized Multimedia Event Processing Scenario

efficient in analyzing and predicting traffic events. Detection of interesting events in sports video [7, 33] is also one of the common event recognition problem. Similarly other applications like flood detection, surveillance based systems, cultural events, and/or natural disasters, are also introduced in literature [1, 31, 43, 50] with medium to high precision and no possibility for domain adaptation. It can be concluded that although these event recognition systems achieve high performance, they have no support for large vocabulary which limits their user interface, they also demonstrate the need to merge event based systems with multimedia methods each time the domain changes, and therefore do not support domain adaptation by themselves.

2.2 Domain Adaptive Event Recognition

As existing approaches of processing multimedia data are domain-specific, the research is moving towards the concept of transfer of knowledge from one domain to another [8, 11, 49]. Domain Adaptation is the ability to utilize the knowledge of old domains to identify unknown domains. The model learns from the source domain consisting of labeled data and from the target domain using unlabeled/labeled data, and in most use-cases, data available in the source domain is much more than the target domain [35]. Many approaches [9, 16, 30, 46] with supervised/unsupervised transfer learning have been proposed for domain adaptation and are mainly focused on generalization ability for increasing accuracy not the overall response time. An event recognition in still images by transferring objects and scene representations has been proposed in work [48], where the correlations of the concepts of object, scene, and events have been investigated. Similarly, large scale domain adaptation based approaches [4, 10, 19, 20, 40] are also introduced particularly for the detection of objects and it is desirable to bring their abilities to the core of multimedia event processing.

3 MULTIMEDIA EVENT PROCESSING

3.1 Problem Formulation

The problem is focused on minimizing the response time for the processing of multimedia events in order to answer user queries consisting of unknown subscriptions (unbounded vocabulary), using an adaptive classifier construction approach while achieving high accuracy. It is primarily based on following two dimensions "Response-Time" and "Unknown Subscriptions":

(i) Response-Time: It can be defined as the difference between the arrival and notification time of subscription processed using specific classifiers. Challenges with response-time in multimedia event processing system include the following two cases:
Case 1: Classifier for subscription available
This case contains subscriptions (like car, dog, bus) which are previously known to the multimedia event processing system, and their classifiers are already present in the model. Here *response-time* will depend only on the testing time while excluding training time.
Case 2: Classifier for subscription not available
This scenario includes subscriptions (like person, truck, traffic_light) for which classifiers are not available and unknown to the system. However by using the similarity of new subscriptions with existing base classifiers, we can further classify the present case as:
(a) Subscriptions require classifiers similar to base classifiers: Consider an example of an unknown subscription "truck", classifier for *truck* can be constructed from existing "bus" classifier. Hence domain adaptation time contributes to response-time.
(b) Subscriptions require classifiers completely different from base classifiers: In such scenario, we assume no base classifiers are similar to incoming subscription and response-time must includes cost of training from scratch.

(ii) Unknown Subscriptions: This dimension concerns the ability to recognize new subscriptions with the naming of objects that may not belong to the limited vocabulary of system. The lack of support for unbounded vocabularies is a bottleneck for emerging applications [52], which we are referring to as Unknown Subscriptions.

3.2 Adaptive Multimedia Event Processing

A functional model has been designed for the adaptive multimedia event processing engine (shown in Fig. 2), consisting of various models discussed below:

Event Matcher analyzes user subscriptions (such as *bus*, *car*, *dog*) and image events, and is responsible for the detection of conditions in image events as specified by user query and preparation of notifications that need to be forwarded to users.

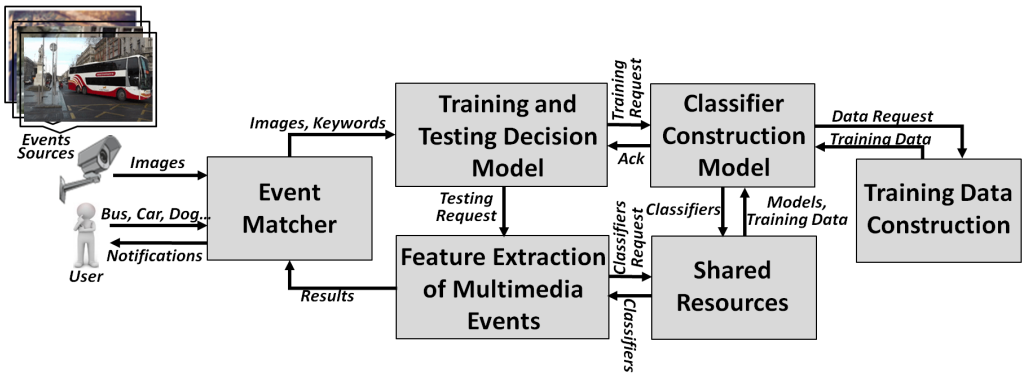


Figure 2: Design for Adaptive Multimedia Event Processing

Training and Testing Decision Model designed to analyze available classifiers and take the *testing* and *training* decision accordingly. It evaluates the relationship of existing classifiers with new/unknown subscription and chooses the *transfer learning* technique.

Classifier Construction Model phase performs the training of classifiers for subscribed classes, and updates the *classifier* in the shared resources after allowed *response-time*. The two options of transfer learning used for classifier construction includes fine-tuning and freezing layers. In the first approach we are performing fine-tuning on a pre-trained model (presently ImageNet [13]), which uses the technique of back-propagation with labels for target domain until validation loss starts to increase. In the second approach, we are using this previously trained classifier to instantiate the network of another classifier required for a similar subscription concept. In this particular scenario, we are freezing the backbone (convolutional and pooling layers) of the neural network and training only top dense fully connected layers, where the frozen backbone is not updated during back-propagation and only fine-tuned layers are getting updated and retrained during the training of classifier. The decision of construction of a classifier for "bus" either from pre-trained models (by fine-tuning) or from "car" classifier (by freezing) is taken with the help of computation of a threshold based on subscriptions relatedness (*path* operator of WordNet [36]).

In *Training Data Construction* model, if a subscriber subscribes for a class which is not present in any smaller object detection datasets (Pascal VOC [15], and Microsoft COCO [28]), then a classifier can be constructed by fetching data from datasets (ImageNet [13], and OID [23]) of more classes using online tools like ImageNet-Utils[1] and OIDv4_ToolKit[2]. Another common approach of online training data construction is to use engines like "Google Images" or "Bing Image Search API" to search for class names and download images.

Feature Extraction of Multimedia Events is responsible for the detection of objects in image events using current deep neural network based object detection models and incorporating new classifiers. Here we utilize image classification models [18, 37, 45] in backbone network of object-detection models.

Shared Resources component consist of existing image processing modules and training datasets. We use *You Only Look Once (YOLO), Single shot multibox detector (SSD), and Focal loss based Dense object detection (RetinaNet)* as object detection models [27, 29, 38, 39]. We have some base classifiers trained off-line using established dataset *Pascal VOC* [15], which are used in constructing more classifiers using domain adaptation.

[1]https://github.com/tzutalin/ImageNet_Utils
[2]https://github.com/EscVM/OIDv4_ToolKit

4 EVALUATION

4.1 Performance With/Without Adaptation

The results of mean Average Precision (mAP) for *response time* from 0 to 30 min are shown in Table 1. In the case of arrival of a completely new subscription (Case 2b in Section–3.1), all models are trained from scratch without use of any pre-trained model. Here, RetinaNet performs higher (mAP $\sim$ 0.21) than other models and the SSD300 does not converge without a pre-trained model. The second and third row indicate the performance of proposed model by applying domain adaptation techniques of fine-tuning/freezing (Case 2a in Section–3.1). The recorded frame rates on our resources for YOLOv3, SSD300, and RetinaNet are 114 fps, 21 fps, and 12 fps respectively, where fps represent the number of frames per second. It can be concluded that domain adaptation via *freezing* layers can provide acceptable performance (i.e. accuracy $\sim$ 92.74% with precision $\sim$ 0.50 using YOLOv3 model) in such short training time (30min) as compared to *fine-tuning* of pre-trained model, which is crucial to know before taking the decision of choosing either pre-trained model or nearest classifier.

4.2 Empirical Analysis for Domain Shift

We analyzes *Transfer Loss*, *Accuracy*, and *Distribution Discrepancy* metrics, for domain adaptation. The "transfer loss" has been evaluated on four domain transfers (varying from closely related domains to not related domains), depicted in Fig. 3a. The transfer achieved by YOLOv3 is better than other object detection models in case of *football* to *cricket ball* and *laptop* to *mango* domain transfers. Here, the transfer loss only indicates how well the transfer works on multiple domains, and lower values are desired. However, the best transfer is achieved by RetinaNet model on the transfer of *cat* to *dog* class. Similarly the single shot detection (SSD) model achieve its best on transfer of *car* to *bus*. Interestingly, the values of transfer loss using models (SSD and RetinaNet) on other domain transfers

Table 1: mean Average Precision (mAP) on Initial ($\alpha = 0$) and Final ($\beta = 30min$) Response-Time With/Without Adaptation

Training Method	YOLOv3		SSD300		RetinaNet	
	α	β	α	β	α	β
Training from Scratch	0.01	0.07	0.00	0.00	0.09	0.21
Fine-Tuning ImageNet	0.00	0.12	0.05	0.17	0.27	0.36
Freezing Similar Classifier	0.15	0.50	0.14	0.16	0.17	0.17

are quit high, and lead us to evaluate *accuracy* on these domain adaptations.

The accuracy achieved by object detection models on the same classes of domain transfers, is shown in Fig. 3b. It can be clearly seen that all object detection models are able to provide high accuracy on applying transfer learning techniques, however the YOLOv3 achieve the best accuracy on all domain transfers.

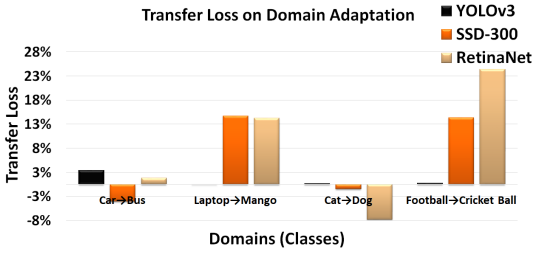
In-order to realize the variation of approximate distance (i.e. Distribution Discrepancy) among different domains, we have trained few binary classifiers that can classify source-target pair of classes like *cat and dog*, *car and bus* etc. It can be seen in the results (Fig. 3c), that distribution discrepancy (lower is better) for YOLOv3 is relatively smaller among most of the domain transfers than for other object detection models, which suggests that YOLOv3 neural network closes the cross-domain gap more effectively, which also explains its better accuracy than other object detection models.

4.3 Evaluations on Known/Unknown Domains

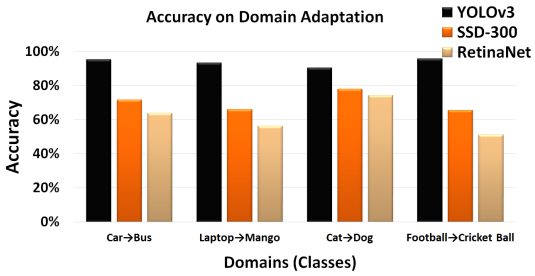
As results of high performance and domain shifts are in favor of YOLOv3 with freezing layer based transfer learning technique, we have selected YOLOv3 as an object detection model for performing further experiments on the unknown subscriptions. Table 2 provides a comparison of average accuracy and response time of Adaptive Multimedia Event Processing model with existing domain-specific models by considering their best performance. It can be observed that existing multimedia event recognition models are designed only for the detection of specific objects and answer such known subscriptions in low response time, while fails to process any unknown subscription. An average response time of the approach for known subscriptions depends only on testing time (~0.01 min) and accuracy (98.53%) of object detection model. However, response time for an unknown subscription includes training (presently ~30 min) via domain adaptation and achieves the accuracy of 95.14%.

5 CONCLUSION AND FUTURE WORK

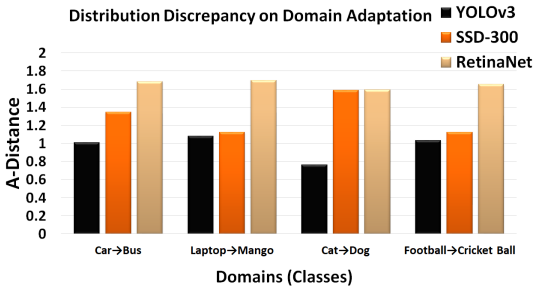
This paper analyzed the problem of processing multimedia events (specifically object detection), for known/unknown subscriptions/concepts, while minimizing the response time. We proposed a multimedia event processing model with domain adaptation by utilizing transfer learning based techniques (fine-tuning and freezing), for the online training of neural network based models. Experiments on current models evaluated the performance in low response-time, along with an empirical analysis for domain shift. The proposed system can achieve accuracy ranges from 95.14% to 98.53% within ~ 0.01 min to ~ 30 min of response-time using YOLOv3 even when subscriptions are unknown. In future work, it can be extended for unsupervised/semi-supervised learning to reduce the need of labeled data for new subscriptions.



(a) Transfer Loss



(b) Accuracy



(c) A-Distance

Figure 3: Analysis for Domain Shift

Table 2: Comparison of Proposed with Existing Model(s)

Approach	Subscription	Performance	
		Response Time	Accuracy
Vehicle Detection for Traffic [47]	Known	0.001 min	97.3%
	Unknown	∞	0%
Firearm Detection for Security [25]	Known	0.0001 min	94.00%
	Unknown	∞	0%
Stolen Object Detection [41]	Known	0.0007 min	93.58%
	Unknown	∞	0%
Car Parking Vacancy Detection [22]	Known	0.17 min	97.9%
	Unknown	∞	0%
Adaptive Multimedia Event Processing Model	Known	0.01 min	98.53%
	Unknown	29.99 min	95.14%

ACKNOWLEDGMENTS

This work was supported by *Science Foundation Ireland* under grant SFI/12/RC/2289_P2. Titan Xp GPU used was donated by NVIDIA.

REFERENCES

[1] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. 2017. Convolutional Neural Networks for Disaster Images Retrieval.. In *MediaEval*.

[2] Sufyan Almajali, I Dhiah el Diehn, Haythem Bany Salameh, Moussa Ayyash, and Hany Elgala. 2018. A distributed multi-layer MEC-cloud architecture for processing large scale IoT-based multimedia applications. *Multimedia Tools and Applications* (2018), 1–22.

[3] Sheeraz A Alvi, Bilal Afzal, Ghalib A Shah, Luigi Atzori, and Waqar Mahmood. 2015. Internet of multimedia things: Vision and challenges. *Ad Hoc Networks* 33 (2015), 87–111.

[4] Asra Aslam. 2020. Object Detection for Unseen Domains while Reducing Response Time using Knowledge Transfer in Multimedia Event Processing. In *Accepted for Proceedings of the 2020 ACM on International Conference on Multimedia Retrieval (ICMR)*.

[5] Asra Aslam and Edward Curry. 2018. Towards a Generalized Approach for Deep Neural Network Based Event Processing for the Internet of Multimedia Things. *IEEE Access* 6 (2018), 25573–25587.

[6] Asra Aslam, Souleiman Hasan, and Edward Curry. 2017. Challenges with image event processing: Poster. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*. 347–348.

[7] Noboru Babaguchi, Yoshihiko Kawai, and Tadahiro Kitahashi. 2002. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE transactions on Multimedia* 4, 1 (2002), 68–75.

[8] Oscar Beijbom. 2012. Domain adaptations for computer vision applications. *arXiv preprint arXiv:1211.4860* (2012).

[9] Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 17–36.

[10] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3339–3348.

[11] Gabriela Csurka. 2017. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374* (2017).

[12] Gianpaolo Cugola and Alessandro Margara. 2012. Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR)* 44, 3 (2012), 15.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[14] Patrick Th Eugster, Pascal A Felber, Rachid Guerraoui, and Anne-Marie Kermarrec. 2003. The many faces of publish/subscribe. *ACM Computing Surveys (CSUR)* 35, 2 (2003), 114–131.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[17] Holger Glasl, David Schreiber, Nikolaus Viertl, Stephan Veigl, and Gustavo Fernandez. 2008. Video based traffic congestion prediction on an embedded system. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*. IEEE, 950–955.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Judith Hoffman. 2016. *Adaptive learning algorithms for transferable visual recognition*. University of California, Berkeley.

[20] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*. 3536–3544.

[21] Ling Hu and Qiang Ni. 2017. IoT-driven automated object detection algorithm for urban surveillance systems in smart cities. *IEEE Internet of Things Journal* 5, 2 (2017), 747–754.

[22] Jermsak Jermsurawong, Mian Umair Ahsan, Abdulhamid Haidar, Haiwei Dong, and Nikolaos Mavridis. 2012. Car parking vacancy detection and its application in 24-hour statistical analysis. In *2012 10th International Conference on Frontiers of Information Technology*. IEEE, 84–90.

[23] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages* 2 (2017), 3.

[24] Malaram Kumhar, Gaurang Raval, and Vishal Parikh. 2019. Quality Evaluation Model for Multimedia Internet of Things (MIoT) Applications: Challenges and Research Directions. In *International Conference on Internet of Things and Connected Technologies*. Springer, 330–336.

[25] Mikolaj E Kundegorski, Samet Akçay, Michael Devereux, Andre Mouton, and Toby P Breckon. 2016. On using feature descriptors as visual words for object detection within x-ray baggage security screening. (2016).

[26] Ching-Hao Lai and Chia-Chen Yu. 2010. An efficient real-time traffic sign recognition system for intelligent vehicles with smart phones. In *Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on*. IEEE, 195–202.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015).

[31] Laura Lopez-Fuentes, Joost van de Weijer, Marc Bolanos, and Harald Skinnemoen. 2017. Multi-modal Deep Learning Approach for Flood Detection.. In *MediaEval*.

[32] Badri Mohapatra and Prangya Prava Panda. 2019. Machine learning applications to smart city. *ACCENTS Transactions on Image Processing and Computer Vision* 4 (14) (Feb 2019). https://doi.org/10.19101/TIPCV.2018.412004

[33] Pirkko Mustamo. 2018. *Object detection in sports: TensorFlow Object Detection API case study*. University of Oulu.

[34] Ali Nauman, Yazdan Ahmad Qadri, Muhammad Amjad, Yousaf Bin Zikria, Muhammad Khalil Afzal, and Sung Won Kim. 2020. Multimedia Internet of Things: A Comprehensive Survey. *IEEE Access* 8 (2020), 8202–8250.

[35] Sinno Jialin Pan, Qiang Yang, et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[36] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*. Association for Computational Linguistics, 38–41.

[37] Joseph Redmon. 2013–2016. Darknet: Open Source Neural Networks in C. http://pjreddie.com/darknet/.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

[39] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242* (2016).

[40] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *European conference on computer vision*. Springer, 213–226.

[41] Juan Carlos San Miguel and José M Martínez. 2008. Robust unattended and stolen object detection by fusing simple algorithms. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 18–25.

[42] Kah Phooi Seng and Li-Minn Ang. 2018. A Big Data Layered Architecture and Functional Units for the Multimedia Internet of Things (MIoT). *IEEE Transactions on Multi-Scale Computing Systems* (2018).

[43] Chiao-Fe Shu, Arun Hampapur, Max Lu, Lisa Brown, Jonathan Connell, Andrew Senior, and Yingli Tian. 2005. Ibm smart surveillance system (s3): a open and extensible framework for event based surveillance. In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*. IEEE, 318–323.

[44] Javier Silvestre-Blanes, Víctor Sempere-Payá, and Teresa Albero-Albero. 2020. Smart Sensor Architectures for Multimedia Sensing in IoMT. *Sensors* 20, 5 (2020), 1400.

[45] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation.. In *AAAI*, Vol. 6. 8.

[47] Yong Tang, Congzhe Zhang, Renshu Gu, Peng Li, and Bin Yang. 2017. Vehicle detection and recognition for intelligent traffic surveillance system. *Multimedia tools and applications* 76, 4 (2017), 5817–5832.

[48] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. 2018. Transferring deep object and scene representations for event recognition in still images. *International Journal of Computer Vision* 126, 2-4 (2018), 390–409.

[49] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[50] Xiu-Shen Wei, Bin-Bin Gao, and Jianxin Wu. 2015. Deep spatial pyramid ensemble for cultural event recognition. In *Proceedings of the IEEE international conference on computer vision workshops*. 38–44.

[51] Piyush Yadav and Edward Curry. 2019. VidCEP: Complex Event Processing Framework to Detect Spatiotemporal Patterns in Video Streams. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2513–2522.

[52] Yuhao Zhang and Arun Kumar. 2019. Panorama: a data system for unbounded vocabulary querying over video. *Proceedings of the VLDB Endowment* 13, 4 (2019), 477–491.