Object Recognition and Tracking for Smart Audio Guides

Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio,
Andrea Ferracani, Marco Bertini and Alberto Del Bimbo

University of Florence, Italy
{name.lastname}@unifi.it

Abstract. In this paper we address the problem of creating a smart audio guide that adapts to the actions and interests of tourists. Our guide performs automatic recognition of artworks and allows the users instant or deferred fruition of multimedia content. We use a compact CNN as computer vision system to back the whole application to performs object classification, localization and recognition. Tracking is used to improve the recognition accuracy over sequences of detections. We also provide an automatic pipeline for dataset creation based on the same tracking algorithm. The system, deployed on an NVIDIA Jetson TK1 and an NVIDIA Shield Tablet, has been tested in a real world environment.

Keywords: Object Recognition, Cultural Heritage

1 Introduction

According to recent statistics from the US National Travel and Tourism Office, a new record of tourism-related activities [1] has been set recently. Museum visits are rising steadily thanks to the availability of new digital and mobile technologies. Modern visitors do not follow fixed paths, but they expect personalization and interaction. As a result, new companion tools are needed, providing content sized to the needs and interests of the visitors [1].

In order to automatically gather the behavior of users, these tools have been using cameras to observe where the users go and what they observe. Several approaches resorted to computer vision systems to offer recommendation based on passive external behaviour observation [2] or, more recently, to develop a wearable smart audio guide [3] to automatically play content or interact with artworks [4]. These modern approaches work by constantly matching the user point of view with a visual database of the known artworks, deciding, depending on user behavior, whether to start or not the audio description generated by means of text to speech technology [3] or to show additional content on gestures [4]. Although designed to work in different settings, they all require a computer vision expert to train and test a computer vision models for artworks, person

[1] http://tinet.ita.doc.gov/tinews/archive/tinews2017/20170413.asp

or statues. Moreover, every time an artwork is added or removed, the database has to be updated and a new model has to be trained. We argue that a more efficient solution would be to let the museum curator add new artworks by himself, without requiring to retrain the model from scratch.

In this paper we propose a wearable audio guide system that, by observing in real time what the user is looking at and by following him in the visit, provides personalized content when needed. The device observes the wearer behavior through a computer vision system and decides when to start and stop the reproduction of the audio content. In contrast to previous work [3], artworks are recognized from an on-board database that can be easily made by museum curators using a novel, easy to use procedure. To this end, we show how to avoid re-training the object detector by learning a generic artwork detector based on convolutional neural network (CNN). We develop a novel artwork tracking technique that is used to populate the database using the same CNN object detector we trained for recognition.

We implemented an Android application that a museum curator can use to build a dataset of artworks adaptively. After a recording phase, it performs all required computation on board and outputs ready to use models for in smart audio guides like [3].

2 Related Work

Our work is mainly related to the personalization of the cultural experience and content recommendation on mobile devices. Many works propose to use mobile systems to enjoy an augmented personalized experience on cultural heritage. One of the first concept was that of Abowd *et al.* [5], that marked the difference between indoor and outdoor systems. We thus follow that division and differentiate between local systems to be used in cultural heritage sites, where there is control over the artworks, and outdoor systems that can be used while traveling in a city.

Local systems are mostly developed for museums. In [6] the Cultural Heritage Information Personalization (CHIP) system was proposed, where a personalized tour could be created through a web interface. The tour can be downloaded to a mobile device using RFID present in the museum, and keeps track of the visited artwork on the server side user profile for successive tours. Analyzing and predicting the behavioral patterns of museum visitors, through the use of interactive digital guides was proposed in [7] and [8]. They follow the four identified patterns, emerged through ethnographic observations in [9]. Augmented reality on a mobile device was explored in [10] to offer a personalized interactive storytelling experience. Based on the age of the visitor, the system provides a gamified experience to children. In [2] a non-intrusive computer-vision system has been employed to perform re-identification and tracking of users in a museum. By observing the interest of the visitors, it can build a user profile that is then used to create a personalized exploration of multimedia content on an interactive table.

Differently from all of these works, we developed a wearable agent that observes the same scene as the user and provides a novel contextually aware interaction based on audio only, that is unintrusive. Moreover, all the computation required is performed onboard.

3 Efficient Object Detection and Recognition

The smart audio guide we developed is based on an efficient computer vision pipeline that simultaneously performs artwork localization and recognition. The guide requires two main computer vision tasks to be solved: *i)* detection of relevant object categories: e.g. persons and artworks; and *ii)* for every detected artwork, reliable recognition of the specific artwork framed. Moreover, since we are dealing with a sequence of frames, in order to improve artwork recognition we take advantage from temporal coherence to make the output more stable. Our method is based on [3], which we briefly cover in the following. We use YOLO [11] network that has the main advantage of processing each frame just once to locate all the objects of interest yielding accurate results even for moderate size networks. The architecture is derived from *Tiny Net*, a small CNN pre-trained on ImageNet, which allows the application to run at 10 FPS and fitting on the memory of a Shield Tablet. The system was fine-tuned to recognize artworks and people using our dataset. Recognizing people is relevant for two reasons: first we can exploit the presence of people in the field of view to create a better understanding of context; secondly, without learning a person model it is hard to avoid false positives on people, since artwork training data contains statues which may picture human figures. Learning jointly a person and an artwork model, the network features can be trained to discriminate between this two classes.

3.1 Artwork recognition

The rich features computed by the convolutional layers are exploited and reused to compute an object descriptor for artwork recognition. To obtain a low dimensional fixed size descriptor of a region, we apply a global max-pooling over two convolutional feature activation maps and concatenate the result, as shown in Fig. 1. The region is remapped from the frame to the convolutional activation map with a simple similarity transformation. As also detailed in [3], through experimental evaluation, we selected the features from layers 3 and 4, yielding a feature of size 768.

Considering a pre-acquired dataset of artwork patches $p_i \in \mathcal{D}$ and their artwork labels y, for each detected artwork d we predict a specific artwork label $y_{\hat{p}}$ finding the nearest neighbor patch

$$\hat{p} = \arg\max_i \langle p_i, d \rangle \tag{1}$$

The recognition system observes each frame independently and predicts artwork labels according to Eq. 1, this approach, in case of motion blur or quick lighting changes may produce incorrect recognition results.

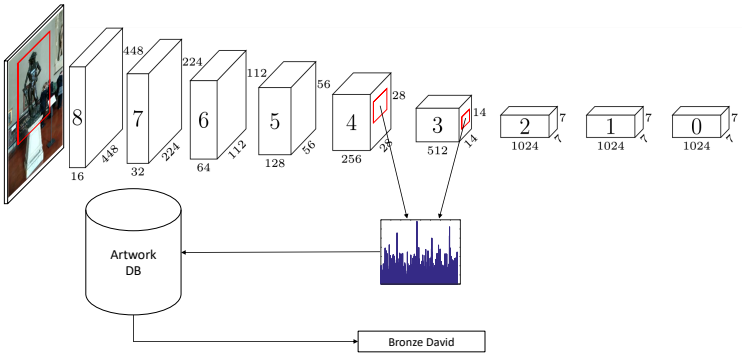

Fig. 1: Our pipeline for recognition, showing network architecture, feature pooling.

4 Automatic Dataset Creation

Extending the dataset with our architecture is extremely straightforward. We rely on a simple multi-target tracking algorithm. With respect to [3] we added a functionality to manage two new use cases: *i) adding a new artwork*, which is needed in the deployment phase of our system to populate the Artwork DB and whenever a new piece is added to the exhibition; and *ii) adding examples of an existing artwork*, which arise at any time the position of artworks or any other environmental condition has caused a decrease in performance of the recognition. Moreover, this second use case allows the exhibition curator to acquire artwork samples at multiple times making the acquisition process easier and less fatiguing.

We perform tracking by data association, first we detect all artworks using our CNN, then we greedily associate bounding boxes to the one detected on the previous frame, allowing association only if intersection over union is above 60%. All unassociated boxes are stored and an association is attempted with all boxes at the following frame. All boxes from the previous frame which could not be associated are removed. This method may fail in case the detector skips a frame, nonetheless we found out that this is a very infrequent case and we allow the user to re-initialize the tracker in case the tracked object is lost.

We only retrieve features for an artwork at a time. When the acquisition view is started, the user is prompted to select one of the detected artworks, enclosed in a dashed bounding boxes as show in Fig. 2, from the rolling video. Once an artwork is selected with a tap, its bounding box is drawn with a solid line and a tracking id is shown (for debugging purposes). For every associated detection our CV system stores in the App database the feature extracted using the method described in Sec. 3 together with the relative frame snapshots

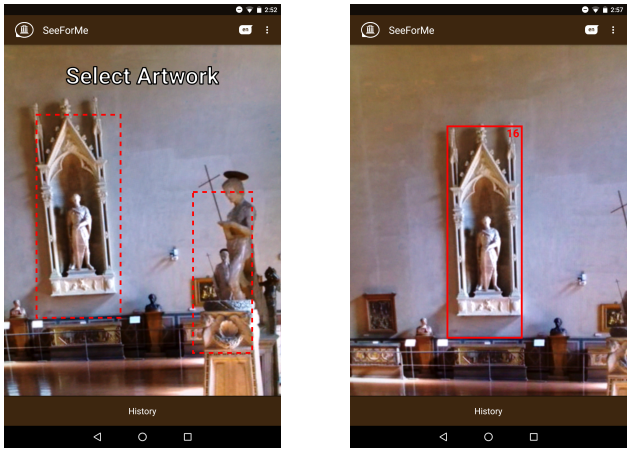

Fig. 2: Tracking process example. User initialization is requested showing all available objects with dashed contours (left). Once tracking has started the tracked object bounding box is shown with solid line and its tracking id (right).

5 Experiments

To show the benefits of our approach for dataset extension in our system, we conduct a simple experiment. We progressively increase the dataset of our artworks reaching a maximum of roughly 2k samples for eight artworks. It can be seen in Fig. 3 that recognition accuracy increases with the amount of samples. It has to be noted that just with the 10% of our acquisition we can reach more than 90% accuracy. Nonetheless increasing the samples reaches almost 100% accuracy. As also shown in [3] the 1-NN approach is best.

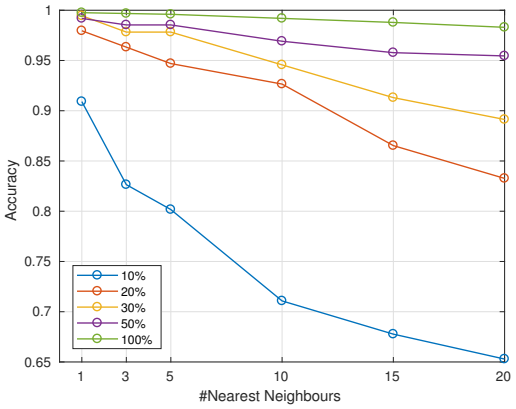

Fig. 3: Accuracy of recognition increasing the dataset size.

6 Conclusions

We have presented a mobile application able to deliver real-time audio information. The main issue of computer vision systems is the training and deployment. We avoid re-training the object detector by learning a generic artwork detector. We show how to populate the database using the same CNN object detector we trained for recognition. Experiments show the benefit of increasing the samples in our pipeline. The system have been deployed and tested on a NVIDIA Shield with TK1.

References

1. Bowen, J.P., Filippini-Fantoni, S.: Personalization and the web from a museum perspective. In: Proc. of Museums and the Web (MW). (2004)
2. Karaman, S., Bagdanov, A.D., Landucci, L., D'Amico, G., Ferracani, A., Pezzatini, D., Del Bimbo, A.: Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. Multimedia Tools and Applications **75**(7) (2016) 3787–3811
3. Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., Bimbo, A.D.: Deep artwork detection and retrieval for automatic context-aware audio guides. ACM Trans. Multimedia Comput. Commun. Appl. **13**(3s) (June 2017) 35:1–35:21
4. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition using wearable vision sensors to enhance visitors museum experiences. IEEE Sensors Journal **15**(5) (2015) 2705–2714
5. Abowd, G.D., Atkeson, C.G., Hong, J., Long, S., Kooper, R., Pinkerton, M.: Cyberguide: A mobile context-aware tour guide. Wireless networks **3**(5) (1997) 421–433
6. Wang, Y., Stash, N., Sambeek, R., Schuurmans, Y., Aroyo, L., Schreiber, G., Gorgels, P.: Cultivating personalized museum tours online and on-site. Interdisciplinary Science Reviews **34**(2-3) (2009) 139–153
7. Zancanaro, M., Kuflik, T., Boger, Z., Goren-Bar, D., Goldwasser, D.: Analyzing museum visitors' behavior patterns. In: Proc. of International Conference User Modeling (UM). (2007)
8. Kuflik, T., Boger, Z., Zancanaro, M. In: Analysis and Prediction of Museum Visitors' Behavioral Pattern Types. Springer Berlin Heidelberg (2012) 161–176
9. Eliseo, V., Martine, L.: Ethnographie de l'exposition. Études et recherche, Centre Georges Pompidou, Bibliothèque publique d'information (1991)
10. Keil, J., Pujol, L., Roussou, M., Engelke, T., Schmitt, M., Bockholt, U., Eleftheratou, S.: A digital look at physical museum exhibits: Designing personalized stories with handheld augmented reality in museums. In: Proc. of Digital Heritage International Congress (DigitalHeritage). (2013)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR). (2016)