# Real-time Wearable Computer Vision System for Improved Museum Experience

Giovanni Taverriti*, Stefano Lombini*, Lorenzo Seidenari,
Marco Bertini and Alberto Del Bimbo
Media Integration and Communication Center, Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy
{*name.surname*}@unifi.it

## ABSTRACT

The goal of this demo is to perform real-time object classification and artwork recognition using a wearable device, to improve user experience during a museum visit by providing contextual information and performing user profiling. We propose the use of a compact CNN network that performs object classification and artwork localization and, using the same CNN features, we perform a robust artwork recognition. Shape based filtering, artwork tracking and temporal filtering further improve recognition accuracy.

## Keywords

Object recognition; Context recognition; cultural heritage

## 1. INTRODUCTION

The goal of this work is to implement a real-time computer vision system that can run on wearable devices to perform object classification and artwork recognition, to improve the experience of a museum visit through understanding the interests of users. Object classification helps to understand the context of the visit, e.g. differentiating when a visitor is talking with people, or just wandering through the museum, or if he is looking at an exhibit that interests him. Artwork recognition allows to provide automatically information of the observed item or to create a user profile based on what and how long a user has observed artworks.

## 2. THE SYSTEM

We base our wearable computer vision system on the YOLO [1] algorithm. To adhere to the real-time requirement on a NVIDIA Jetson TK1 board, we use an architecture derived from the *Tiny Net* proposed by authors of [1].

YOLO has an interesting structure since it combines recent advancements in deep convolutional neural network design with the novel idea of dealing with object detection as a regression problem.

*Equal contribution, {*name.surname*}@stud.unifi.it



*MM '16 October 15-19, 2016, Amsterdam, Netherlands*



ACM ISBN 978-1-4503-3603-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2964284.2973813

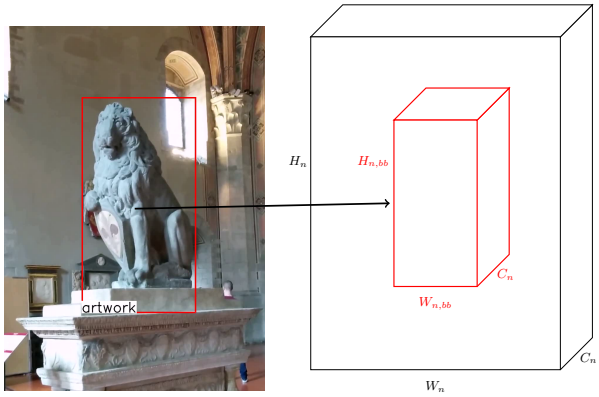


**Figure 1: Feature extraction procedure for an artwork detection on a single convolutional feature map.**

The network is designed taking inspiration from GoogleNet inception modules, which exploits $1\times1$ convolution to reduce the dimensionality and the amount of computation.

Given a convolutional feature map, computed after a set of non-linear transformations, the final Fully Connected (FC) layer for each of the evaluated windows generate a bounding box and a vector of class probabilities. After remapping this $|C| \times |B| \times 5$ tensor to a set of windows, non-maximal suppression is applied to get rid of redundant detections.

We fine-tuned the network to recognize artworks and people using our publicly available dataset[1]. The CNN features computed for classification are then used to recognize specific artworks, e.g. to provide informations through an automatic audio guide, and to understand users behaviours and interests, based on how long an artwork is observed.

### 2.1 Artwork recognition

To obtain a lightweight but descriptive visual feature to perform artwork recognition we re-map artwork detections onto convolutional feature maps and apply a global max-pooling (Fig. 1). Consider our network architecture in Fig. 2. Layers on the left are high-resolution and very correlated with low-level image features, while layers on the right end are low-resolution, highly semantic representation of the image content. We use feature maps 3 and 4, for a descriptor of size 768, obtained concatenating the max-pooled tensors extracted from the rescaled regions.

[1]https://www.micc.unifi.it/resources/

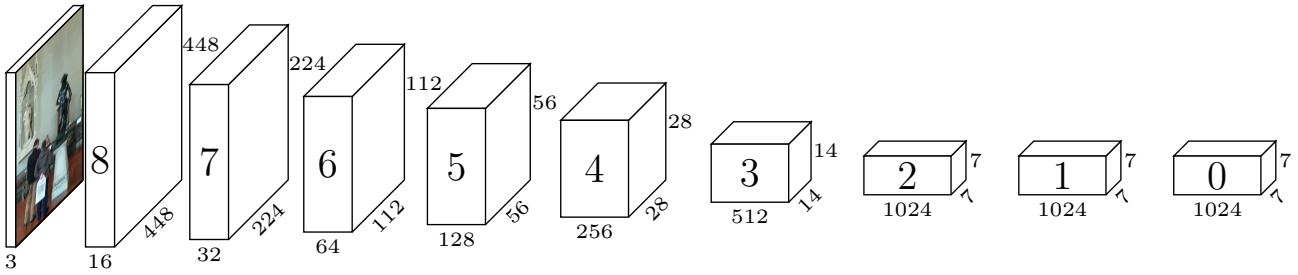


**Figure 2: Our network architecture, with tensor size and layer numbering.**

Considering a dataset of artwork patches $p_i \in \mathcal{D}$ and their artwork labels $y$, we annotated each artwork detection $d$ finding the nearest neighbor patch

$$\hat{p} = \arg\max_i \langle p_i, d \rangle \quad (1)$$

and associating the respective label $\hat{y}$. According to our experimental evaluation layer 3 was the best performing alone. The coupling with layer 4 gave the best result.

We aim at high recognition accuracy, since mistaking an artwork for another may result in a bad user experience, e.g. due to the audio guide that presents an artwork different from the one that is actually observed.

We avoid recognizing detected artwork which are evaluated to be too distant to be of interest. Not having any camera calibration we rely on a simple heuristic comparing the artwork bounding box area with the frame area:

$$\frac{w_{bb} h_{bb}}{WH} > \alpha \quad (2)$$

where $WH$ is the frame area and $w_{bb}$ and $h_{bb}$ are bounding box width and height respectively, and $\alpha$ is a threshold (Fig. 3). In our experiments we obtained the best results for $\alpha = 0.05$, that allows to reduce false recognitions by 50% w.r.t. not using the heuristic, at the cost of introducing a small number of missed recognitions.

## 2.2 Temporal Smoothing

We apply two different strategies to improve the stability of our recognition.

We continuously predict artwork labels according to Eq. 1, but we consider a prediction only after it persist for $M$ frames. We implement this by tracking all artwork detection boxes with a greedy data association tracking-by-detection algorithm.

To further improve recognition result we apply the following strategy. We increment a counter $p$ every time the recognition label for a box is unchanged, keeping track of the most frequent label $\overline{y}$. Every time a label $y*$ different from $\overline{y}$ we decrement $p$. We predict the artwork identity as $y*$ only if $p > P > M$. This technique greatly reduces the number of false recognitions. In our experiments best results were obtained for $M = 15$ and $P = 20$.

## 2.3 Dataset

We collected people images from PASCAL VOC2007 and an in house dataset comprising images from museums. Artwork images have been collected shooting videos inside the Bargello Museum. 8 masterpieces of Donatello have been selected as artworks to be recognized. A tool to easily add artwork descriptors from videos has also been developed, to ease development in other museums.

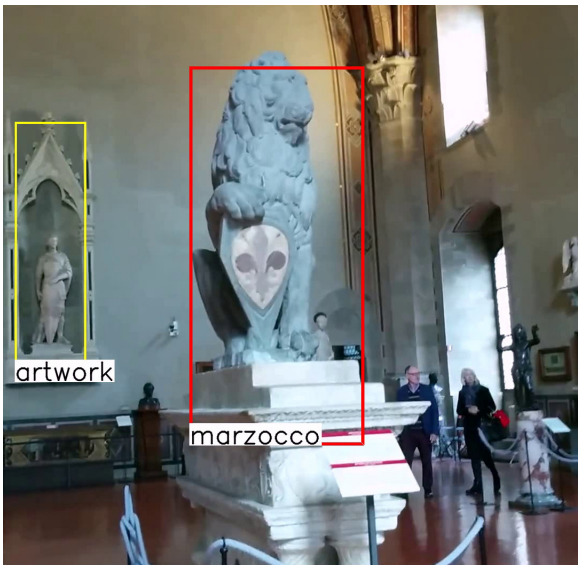


**Figure 3: Shape based filtering: artwork in yellow (left) is not considered for recognition, not satisfying Eq. 2, while the other is recognized as "marzocco".**

# 3. CONCLUSIONS

We have presented a system running on the NVIDIA Jetson TK1 SoC. Our approach jointly solves two problems: contextual analysis and object recognition. We apply our efficient pipeline to improve museum experience. Our method allows to profile in real-time visitor interest and to provide instantaneous feedback on the artworks of interest.

*Acknowledgments.*

This work is partially supported by the "Social Museum and Smart Tourism" project (CTN01_00034_231545).

# 4. REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.