From a Still Image to a Semantically Aware Video: A Context and Metadata-driven Automatic Media Production Framework

Paula Viana *†, Pedro Carvalho †*
{paula.viana,pedro.carvalho}@inesctec.pt
Maria Teresa Andrade ‡†, Inês N. Teixeira *†‡
{maria.t.andrade,ines.f.teixeira}@inesctec.pt
Pieter P. Jonker §
p.p.jonker@qdepq.com
Luís Vilaça ‡†*, José Pedro Pinto †, Tiago Costa †‡
{luis.m.salgado,jose.p.pinto,tiago.a.costa}@inesctec.pt

* Polytechnic of Porto
Porto, Portugal
† INESC TEC
Porto, Portugal
‡ University of Porto
Porto, Portugal
§ QdepQ Systems
Delft, The Netherlands

Due to recent technology outbreaks, anyone can become a mass content producer, spreading multimedia content widely and at a high speed in the web. The amount and variety of photos and videos available can be overwhelming, thus bringing new challenges to the marketing industry, specifically in re-purposing content for advertisement and publishing promotional content rapidly. As such, it is imperative to find supporting automatic systems for enabling an immediate production of engaging marketing content. FotoinMotion proposes an innovative, fast, and cost-effective solution using static media, i.e. content and context information, for the creation of short storytelling videos, using a still photograph as baseline.

For building these rich contextualized multimedia stories, our framework brings together distinct automated tools for contextual data extraction, object recognition, creative transformation, 3D-editing, and text animation, to produce highly engaging experiences. Experiments with users have been currently focusing on three major creative industries: (1) photojournalism, to develop immersive on-filed photo-driven stories, (2) fashion, to create new forms of marketing, product placement and event coverage, (3) festivals, to enable public relations and publicity managers to communicate festival experiences and engage audiences through immersive communication. Nevertheless, the solution can be extended to other creative industries and scenarios.

Creating engaging videos for promotional purposes built from a picture has been a common topic in the literature. There are several commercial solutions for producing low cost videos using predefined animations, but they usually do not consider image content information. Thus, common content labelling frameworks could be used for providing content information on images, but the produced labels are assigned to the whole image, and not regions, making them unfeasible for creating object-based animations. Additionally, relevant information for creating automatic summaries can also be obtained by extracting context information using sensors from a mobile phone [2, 3, 4, 6]. Nevertheless, this data is not currently being used to infer extra information or fused with content-based information. Thus, existing solutions contribute for content re-purposing, but even by merging them into a single system, they could never create automatic, customisable, context and content aware at object basis solutions.

Taking the abovementioned goals into consideration, it is viable to build the proposed system following a simple workflow: taking a photo, using a web-based platform or a mobile app, collecting contextual information, using the image as an input for object detection, and building animations based on the collected information. As such, we have implemented a set of modules illustrated in Figure 1, and summed-up below:

- eCAT: acquires context information at the instance the photograph is taken. This information is collected from sensors, by using the Intel Context Sensing Package. By searching external sources, we infer higher level information to find analogous known events and locations.

- iCAT: intelligent algorithms trained for identifying regions of interest in the photograph. Considering user requirements, we have selected specific classes of objects to be identified, i.e. people, clothing items, fashion accessories and symbols, built a dataset of 1500 images, and used transfer learning from Inception-Resnet-v2 [5] and Resnet-101 [1]. We have also built algorithms targeting potential perceptually relevant regions by detecting differences in colour and luminance balance.

- AAT: assisted annotation tool to refine the automatically generated information, that also enables adding annotations manually.

- Anims: a module dedicated to creating engaging and immersive video animations, including effects and filters. Filters modify static content, e.g. brightness or colour. Effects are modifications that change over time, e.g. pan-effect, rotations of objects, zooming in and out. By extracting depth information from the original photograph, motion can be specified in the three-dimensional space, e.g. motion-blur, bokeh and vertigo effects.

- API: module for management and communication

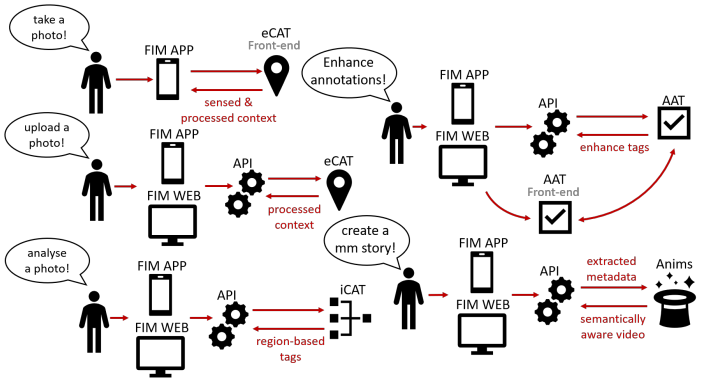

Figure 1: Summed up representation of system interaction, in modules

FotoinMotion contributes to the state-of-the-art in computer vision and multimedia technologies fields and is currently assisting content creators with a high-value-added service, for building rich promotional semantically aware multimedia stories. Our users recognise this system's value and usefulness, acknowledging the positive impact such tool would have on their daily activities. Future work includes improvement or addition of models for object detection and high-level concept identification, techniques for using small training datasets and metadata fusion to increase the knowledge inferred.

The work presented was developed within the project FotoInMotion supported by the European Commission under the contract H2020-ICT-20-2017-1-RIA-780612. We would like to thank the professionals from the different targeted industries for their active participation.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition*, 2016.

[2] Yukun Li, Ming Geng, Fenglian Liu, and Degan Zhang. Visualization of photo album: selecting a representative photo of a specific event. In *Int. Conf. Database Systems for Advanced Applications*, 2019.

[3] Xingjia Pan, Fan Tang, Weiming Dong, Chongyang Ma, Yiping Meng, Feiyue Huang, Tong-Yee Lee, and Changsheng Xu. Content-based visual summarization for image collections. *IEEE Trans. on Visualization and Computer Graphics*, 2019.

[4] Anurag Singh, Lakshay Virmani, and AV Subramanyam. Image corpus representative summarization. In *2019 IEEE 5th Int. Conf. on Multimedia Big Data (BigMM)*, 2019.

[5] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv:1602.07261*, 2016.

[6] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *AAAI*, 2018.