Semantic Storytelling Automation: A Context-Aware and Metadata-Driven Approach

Paula Viana†
Polytechnic of Porto, INESC TEC
Porto, Portugal
paula.viana@inesctec.pt

Pedro Carvalho
INESC TEC, Polytechnic of Porto
Porto, Portugal
pedro.carvalho@inesctec.pt

Maria Teresa Andrade
University of Porto, INESC TEC
Porto, Portugal
maria.t.andrade@inesctec.pt

Pieter P. Jonker
QdepQ Systems, Delft University of Technology
Delft, Netherlands
p.p.jonker@qdepq.com

Vasileios Papanikolaou
Athens Technology Center
Athens, Greece
v.papanikolaou@atc.gr

Inês N. Teixeira
Polytechnic of Porto, INESC TEC, University of Porto
Porto, Portugal
ines.f.teixeira@inesctec.pt

Luís Vilaça
University of Porto, INESC TEC, Polytechnic of Porto
Porto, Portugal
luis.m.salgado@inesctec.pt

José Pedro Pinto
INESC TEC
Porto, Portugal
jose.p.pinto@inesctec.pt

Tiago Costa
INESC TEC, University of Porto
Porto, Portugal
tiago.a.costa@inesctec.pt

ABSTRACT

Multimedia content production is nowadays widespread due to technological advances, namely supported by smartphones and social media. Although the massive amount of media content brings new opportunities to the industry, it also obfuscates the relevance of marketing content, meant to maintain and lure new audiences. This leads to an emergent necessity of producing these kinds of contents as quickly and engagingly as possible. Creating these automatically would decrease both the production costs and time, particularly by using static media for the creation of short storytelling animated clips. We propose an innovative approach that uses context and content information to transform a still photo into an appealing context-aware video clip. Thus, our solution presents a contribution to the state-of-the-art in computer vision and multimedia technologies and assists content creators with a value-added service to automatically build rich contextualized multimedia stories from single photographs.

CCS CONCEPTS

• Computing methodologies ~ Computer vision • Information systems ~ Multimedia content creation

KEYWORDS

Storytelling; Computer Vision; Video Generation; Metadata

†Corresponding author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3416528

ACM Reference format:

Paula Viana, Pedro Carvalho, Maria Teresa Andrade, Pieter P. Jonker, Vasileios Papanikolaou, Inês N. Teixeira, Luis Vilaça, José Pedro Pinto, Tiago Costa. 2020. Semantic Storytelling Automation: A Context-Aware and Metadata-Driven Approach. In *Proceedings of ACM International Conference on Multimedia (MM'20)*, October 12-16, 2020, Seattle, WA, USA, 3 pages. https://doi.org/10.1145/ 3394171.3416528

1 Introduction

The unprecedented growth on available capturing devices and online content brings new opportunities to media professionals but also challenges to guarantee repurposing and producing engaging promotional videos. This work presents an innovative framework for repurposing audio-visual content and automatically produce contextualized animated video stories based on simple photographs.

The work was developed focusing on three main scenarios: photojournalism, fashion, and cinema festivals. Nonetheless, the solution can be adapted to other creative industries or application scopes. The developed framework combines automated tools for context data extraction, object recognition and creative manipulations, merged to create smart immersive experiences. Each one of these modules is described in the following sections.

2 Related Work

Creating engaging videos for promotional purposes taking, as baseline, a picture has been a common topic in the literature. Magisto [1], Animoto [2], Flixel [3] and Faceboook produce low-cost videos using predefined animations, but do not consider

image content information. Content labelling frameworks like Google Vision API [4], Microsoft Vision API [5] and Clarifai [6] provide content information on images, but the produced labels are assigned to the whole image, and not regions. This makes them unviable for creating object-based animations.
Relevant information for creating automatic summaries can also be obtained by extracting context information using sensors from a mobile phone [7][8][9][10]. However, this data is not used to infer extra information or fused with content-based information.
Thus, available solutions contribute for content re-purposing, but do not answer user requirements, for creating automatic, customisable, context aware, at object basis solutions.

3 System Description

Figure 1 depicts the workflow as a user starts by taking a photo using a mobile application or by uploading a photo directly to the main framework. The photos are automatically annotated by intelligent algorithms and this content-based metadata is added to the environment metadata acquired together with the photo. The resulting annotations can be visualized and altered, and the user can add new annotations and write captions with extra information. Finally, the framework creates an automatic story, enhancing the content with filters and effects, using a decision support system based on the collected data. The final output is an animated video produced automatically or customized by the user.

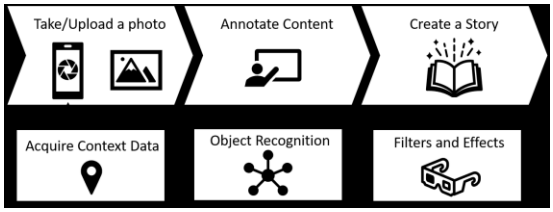

Figure 1: System description – the user experience workflow is presented in arrow-shaped forms, whereas backend modules are illustrated with rectangles

3.1 Upload a photo

By using the mobile app, context information is collected at the moment the photograph is taken. This information is gathered from sensors available on mobile devices, by using the Intel Context Sensing package [11], namely the Activity Recognition, the Audio Classification and the Relative Location. Additionally, by searching external sources, the system automatically infers if the photo is associated with a known event. Alternatively, the user can upload an image in the main framework, without acquiring any context information.

3.2 Annotate Content

Automatic annotations of regions-of-interest (ROI) in the images are produced using computer vision algorithms. Considering user requirements, specific classes of objects to be identified were defined: people, clothing items, fashion accessories and symbols. For this purpose, a dataset with 1500 images, containing over ten thousand objects, was built and transfer learning was applied on previously trained models, specifically Inception-Resnet-v2 [12] and Resnet-101 [13].

Aside from object detection, algorithms targeting potential perceptually relevant regions were used, by detecting differences in colour and luminance balance.
To enhance the automatically generated information, the platform includes an assisted annotation tool that enables adding annotations manually (ROIs and captions) and for creating new datasets for further improvements (e.g. adding a new class).

3.3 Create a story

Engaging and immersive clips are automatically produced from photos using several filters and visual effects. Filters are operations that modify static content, e.g. brightness or colour. Effects are dynamic modifications on the media that change over time. In 2D traditional effects, motion is limited to a bidimensional plane, 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 3D-space, e.g. motion-blur, bokeh and vertigo effects. An example of an automatically generated clip, using the acquired metadata, is shown in Figure 2.

Figure 2: Time sequence illustrating the final animation

4 Conclusions

This paper describes a tool that automatically produces engaging videos from static media. The solution was built on the requirements of media professionals from several sectors, guaranteeing that it will have a positive impact on their activities. Future work includes adding new classes on the object detection module and merging metadata to improve automatic content production.

ACKNOWLEDGMENTS

The work presented in this paper has been supported by the European Commission under contract number H2020-ICT-20-2017-1-RIA-780612.

REFERENCES

[1] Magisto, "Magisto," [Online]. Available: https://www.magisto.com/. [Accessed 01 04 2020]

[2] Animoto, "Animoto," [Online]. Available: https://animoto.com/. [Accessed 01 04 2020].

[3] Flixel, "Flixel," [Online]. Available: https://flixel.com/. [Accessed 01 04 2020].

[4] Google, Google Cloud: Vision AI. [Online]. Available: https://cloud.google.com/vision. [Accessed 18 03 2020].

[5] Microsoft Azure. Microsoft Azure: Computer Vision. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/. [Accessed 18 03 2020].

[6] Clarifai. Clarifai. [Online]. Available: https://www.clarifai.com/. [Accessed 18 03 2020].

[7] J. Wang, J. Fu, J. Tang, Z. Li and T. Mei. 2018. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In Thirty-Second AAAI Conference on Artificial Intelligence.

[8] X. Pan, F. Tang, W. Dong, C. Ma, Y. Meng, F. Huang, T.-Y. Lee and C. Xu. 2019. Content-Based Visual Summarization for Image Collections. IEEE Trans. on Visualization and Computer Graphics.

[9] A. Singh, L. Virmani and A. Subramanyam. 2019. Image Corpus Representative Summarization. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). pp 21-29. https://doi.org/10.1109/BigMM.2019.00-46

[10] Y. Li, M. Geng, F. Liu and D. Zhang. 2019. Visualization of photo album: selecting a representative photo of a specific event. In International Conference on Database Systems for Advanced Applications. Lecture Notes in Computer Science, vol 11448. pp 128-141. https://doi.org/10.1007/978-3-030-18590-9_9

[11] Intel. Intel Context Sensing SDK. 2014. [Online]. Available: https://www.youtube.com/watch?v=DX9wP7ZhAOY. [Accessed 16 03 2020].

[12] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI Conference on Artificial Intelligence. pp. 4278–4284.

[13] K. He, X. Zhang, S. Ren and J. Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778, doi: 10.1109/CVPR.2016.90.