Creating automatic storytelling videos from still images: a semantically-aware approach

Inês N. Teixeira[*†‡], Paula Viana[*†], Maria Teresa Andrade[‡†], Pedro Carvalho[†*], Luís Vilaça[‡†*],
José Pedro Pinto[†], Tiago Costa[†‡], Stamatis Rapanakis[¶], Pieter P. Jonker[§]
[*]Polytechnic of Porto, [†]INESC TEC, [‡]University of Porto. Porto, Portugal
[¶]Athens Technology Center. Athens, Greece, [§]QdepQ Systems. Delft, The Netherlands
{ines.f.teixeira,paula.viana,maria.t.andrade,pedro.carvalho,luis.m.salgado,jose.p.pinto,tiago.a.costa}@inesctec.pt,
s.rapanakis@atc.gr, p.p.jonker@qdepq.com

***Abstract*—In a widely connected world, communicating through multimedia content has exponentially been gaining popularity. Mostly for common use, technology that enables the production of quality content for sharing in social media, plays nowadays an essential role. The main goal is to achieve content production the faster and the better as possible. Yet, taking a photograph and applying a filter is now a quite conventional type of content production. Thus, the new challenge in multimedia communication comes in finding creative approaches for content creation, towards using technology for building new smart and appealing solutions.**

In light of this new market demand, this paper presents Fotoinmotion: a system for producing automatic videos based on a still image. Instead of using animation templates transverse to the content of a photograph, this framework takes into account semantic information to produce storytelling videos, focusing on the relevant features of the input image. The resulting video animations are rich contextualized multimedia stories, presenting content information with visual filters and depth-aware animations.

I. Introduction

Sharing experiences through video is a common practice on social networks. However, with the unprecedented access to multimedia capturing devices, more and more content is shared online, making an overloading amount of information forced into our attention. Although this might bring new challenges for the multimedia technologies community, in terms of organizing and selecting information, it also clouds relevant content shared for promotional purposes. Industries that promote their products through video advertisement suffer from an imminent need for announcing novelty in a valuable and fast way.

By building an intelligent system that could create engaging promotional videos should work as a crutch in this creative process. Such technology could not only be used for professional means, but it could also be an important contribution for common use, in the creation of automatic appealing multimedia content. Fotoinmotion is an European project, with the collaboration of international companies, that aims to create this innovative system, taking into account technological innovation, human interaction and usability, and value-added production in different commercial industries. In a nutshell, this system takes a still image as a baseline, reads relevant content and context information and creates automatic filters and animations, outputting a semantic-aware video, as fast and engagingly as possible.

II. Literature Review

Commercial multimedia platforms and applications are vastly used for sharing videos on social networks. Animoto[1], Magisto[2] and Flixel[3] are tools available online for low-cost production of videos that use standard animations and simple filters, for enhancing the creative content. Nevertheless, these do not consider the actual content of the input picture.

In contrast, in language processing, some relevant approaches have been emerging in the literature for creating narratives automatically, or for photo summarization, basing their solutions on information obtained by mobile devices sensors [4], [5], [6], [7]. Yet, none of these approaches considers inferring relevant high-level information from the collected data.

Frameworks for automatic content annotation, such as Clarifai[8], Google Vision API[9] and Microsoft Vision API[10], are useful for content organization and generic labelling, but they do not enable building automated animations on an object basis.

The above-mentioned approaches contribute for content repurposing, but building fast and low-cost systems for automated video creation, enabling customization, building region-based animations and using storytelling for video sequence, are important aspects that must also be explored.

III. Framework Analysis

Our system is organized in simple task dedicated modules, that will be presented in the next subsections, within a realistic context. The user system interaction tasks are illustrated in 1. They can be summed up in the following steps: uploading a photo into the system, by using a web-based platform or a mobile app, that collects contextual information; running computer vision algorithms for object detection; using a collaborative platform for enhancing automatically produced annotations; using templates or customize features for generating filters and effects, and downloading or sharing the output video.

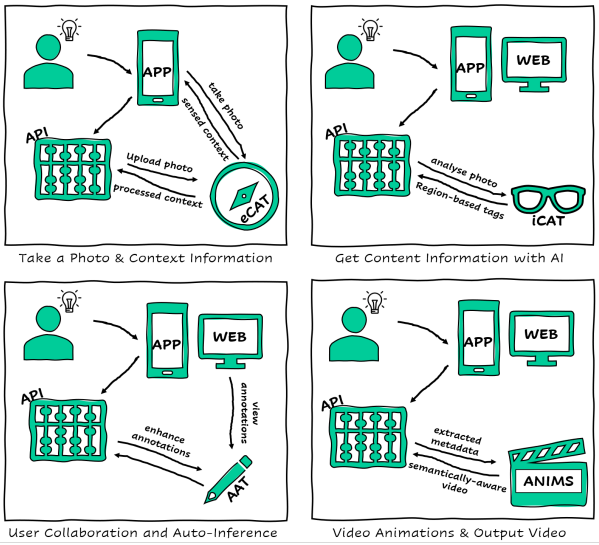

Fig. 1. Summed up representation of system interaction, in modules

A. A Photograph with Context Information (eCAT)

A hired photographer is covering a fashion show for the autumn collection of a clothing brand. The event is about to start, and he used his Fotoinmotion mobile app to take a picture of a model at the entrance. The mobile app uses context information, gathered from sensors information, by using the Intel Context Sensing package[11]. The Relative Location will, later on, be used to infer what was the event happening where the photo was taken. The Audio Classification and Activity recognition can be used for understanding the movement and noise at that moment.

B. Content Information using Intelligent Approaches (iCAT)

The photographer knows that after taking this photo, intelligent algorithms will run to identify relevant content information. Region-based annotations will be produced based on the system known classes: people, clothing items, fashion accessories and symbols. For the Fotoinmotion project, 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]. Some other perceptually relevant regions can also be detected due to the algorithms for identifying saliency regions and high luminance areas. Figure 2 shows the automatically produced results on the left.

C. User Collaboration (AAT)

In order to enhance the generated labels and regions, the photographer uses an assisted annotation platform, which enables the customization of the content annotation process. Figure 2 shows the manually produced results on the right.

D. Automatic Storytelling Animations (Anims)

The brand wishes to produce a video animation quickly, in order to promote the events in social media. The photographer proceeds to request the creation of a video animation.

Fig. 2. Left: obtained annotations after the icat module. Right: user collaboration for annotation enhacing using the aat module

Although the Fotoinmotion platform enables customization, he chooses to apply automatic animations based on the image annotations. The creative tools introduce zooming effects in the identified objects, navigating into the photo in a storytelling sequence, and add visual effects, with color filters and visual depth-aware effects.

IV. Results

The automatic production of promotional videos with engaging and visually appealing strategies was achieved by using collected content and context information, so that the final video is semantically-aware and is built in a storytelling-based sequence. An illustration of an output video storyboard is presented in 3.

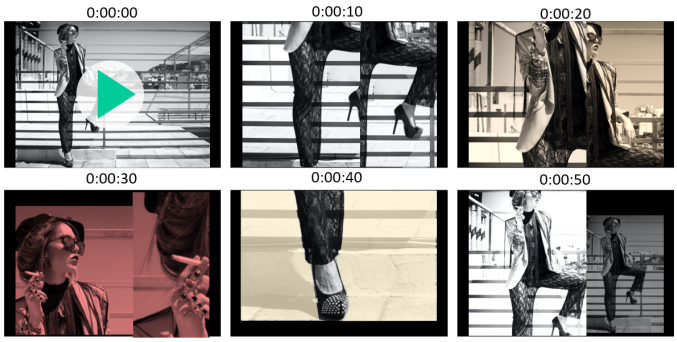

Fig. 3. Storyboard of a Fotoinmotion output video.

V. Conclusion

Testbeds have been conducted with collaborators from creative industries to evaluate our system. The tool had a positive impact on these activities and has been classified as a value-added solution, namely in shortening production time and in creating relevant semantic-aware animations automatically.

It is foreseen as future work to refine the machine learning algorithms for object detection, to reduce the required dataset size for training and to improve high-level inference for concept identification.

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

[2] "Magisto," last accessed: 01/04/2020. [Online]. Available: https://www.magisto.com/

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

[4] J. Wang, J. Fu, J. Tang, Z. Li, and T. Mei, "Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training." in *AAAI*, 2018.

[5] X. Pan, F. Tang, W. Dong, C. Ma, Y. Meng, F. Huang, T.-Y. Lee, and C. Xu, "Content-based visual summarization for image collections," *IEEE Trans. on Visualization and Computer Graphics*, 2019.

[6] A. Singh, L. Virmani, and A. Subramanyam, "Image corpus representative summarization," in *2019 IEEE 5th Int. Conf. on Multimedia Big Data (BigMM)*, 2019.

[7] Y. Li, M. Geng, F. Liu, and D. Zhang, "Visualization of photo album: selecting a representative photo of a specific event," in *Int. Conf. Database Systems for Advanced Applications*, 2019.

[8] "Clarifai," last accessed: 18/03/2020. [Online]. Available: https://www.clarifai.com/

[9] "Google cloud: Vision ai." last accessed: 18/03/2020. [Online]. Available: https://cloud.google.com/vision

[10] "Microsoft azure: Computer vision." last accessed: 18/03/2020. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

[11] Intel, "Intel context sensing sdk," 2014, accessed: 2020-03-16.

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv:1602.07261*, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE conference on computer vision and pattern recognition*, 2016.