BUILDING MODELS FOR MOBILE VIDEO UNDERSTANDING

Franck Ngamkan & Geneviève Patterson
TRASH Inc.
Brooklyn, NY 11201, USA
{franck,gen}@trash.app

EXTENDED ABSTRACT

Social media is constantly raising the bar for content quality. People who shoot video on their phones want easy ways to make their videos look good enough to post on social media. Trimming videos to the best moments, identifying the most cinematically appealing moments, and editing clips together quickly enough to post in the real-life moment are challenging problems for users and researchers who want to automate these tasks.

Computational editing and cinematography are topics of growing interest to the computer vision and graphics communities Leake et al. (2017); Merabti et al. (2016); Truong et al. (2016); Gandhi et al. (2014). Existing video datasets, such as Marszałek et al. (2009); Monfort et al. (2018); Fouhey et al. (2018); Gu et al. (2018); Zhou et al. (2017), lack labels for many cinematic concepts (e.g. wide shot, close-up), do not contain these visual events, or do not contain video captured on mobile devices. We need annotated video that contains examples that have the photographer's bias of our users – portrait aspect ratio, amateur photography, selfies, etc. Existing academic datasets have content and photography biases that are significantly different from the video we are interested in.

To meet our users' needs, our team creates a dataset of cinematic and social concepts in mobile video using a restricted budget of time and capital. In this talk we will address how we (i) recognize visual concepts relevant to our users (ii) train models on source images and video that our users are likely to create (to prevent photographer's bias) (iii) evaluate models trained on weakly or unlabeled data when no labeled test data is available. Finally we demonstrate our improved editing results informed by our social-video event recognition network.

We introduce our strategy to handle an ever-changing landscape of popular visual concepts in the meme era. We present the practical performance of our active learning system, based on our earlier academic research Patterson et al. (2015); Kaspar et al. (2018). This internal tool allows us to bootstrap proprietary datasets of unlabeled user videos with only one or a few labeled starting examples (Fig. 1). Our system has the ability to exploit different neural network feature extractors, learn from video or from still images, and employ a variety of active learning query strategies to quickly learn new visual concepts (Fig. 2). In this talk, we also introduce our custom evaluation metric based on the rank order of correctly identified samples for assessing the performance of a classifier trained in a semi-supervised manner when no ground truth test data is available.

In our interactive demo, attendees will have the opportunity to train new classifiers with our internal active learning tool (screenshot in Fig. 1) and create automatically edited video with the TRASH app (example edit output shown in Fig. 3).

ACKNOWLEDGEMENTS

This work is supported by NSF Phase I SBIR Grant 1842850.

APPENDIX - DEMO FIGURES

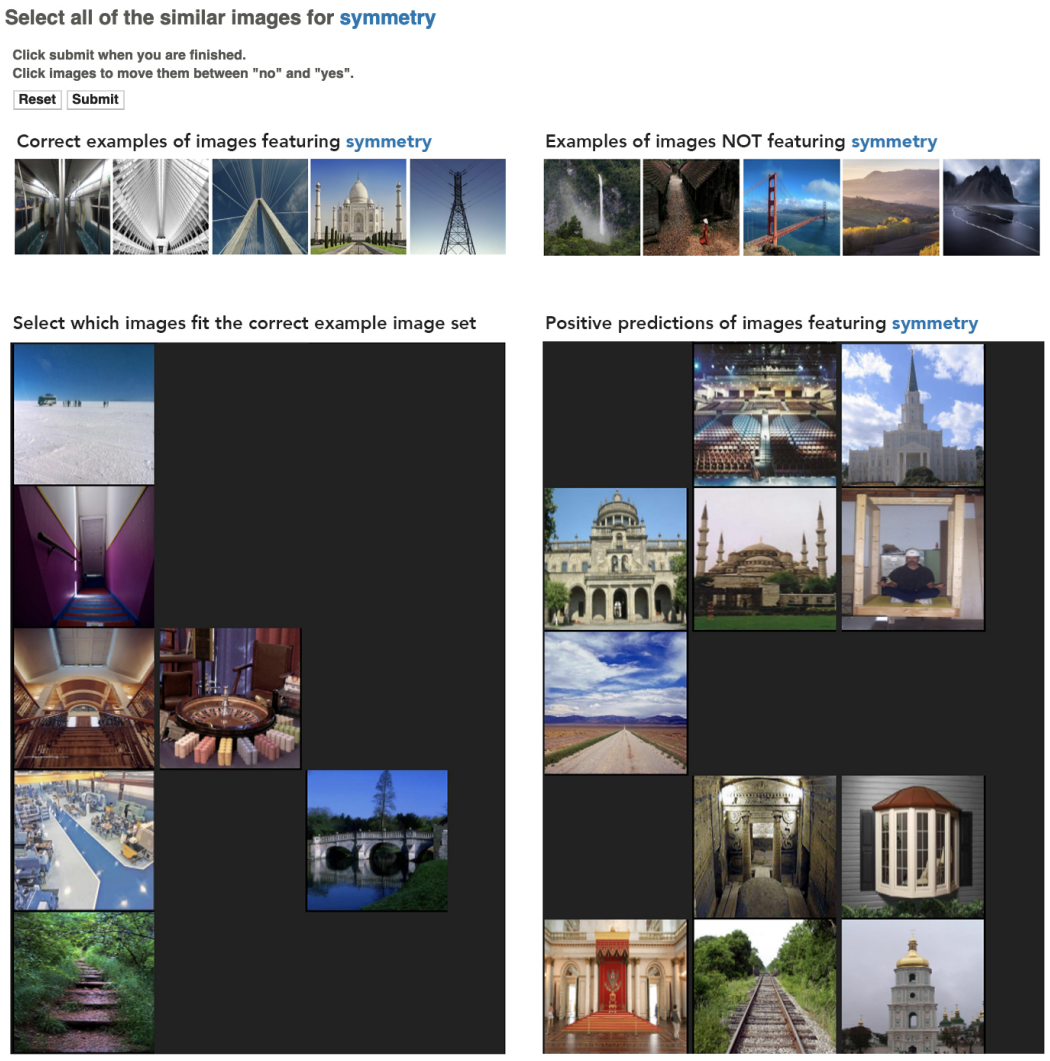

Figure 1: **Video Classifier Bootstrapping UI.** This is annotator UI for responding to active queries in our internal active learning tool. In this figure, the annotator is creating a classifier for the concept *symmetry*.

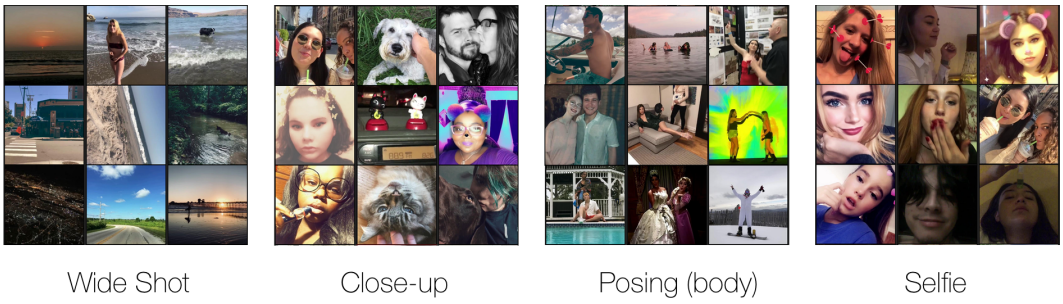

Figure 2: **Results of Successfully Bootstrapped Video Classifiers on Unlabeled Test Set.** These are sample detections of the concepts listed above discovered in an unlabeled test set of user videos. Classifiers were trained using the active learning system in Fig. 1.

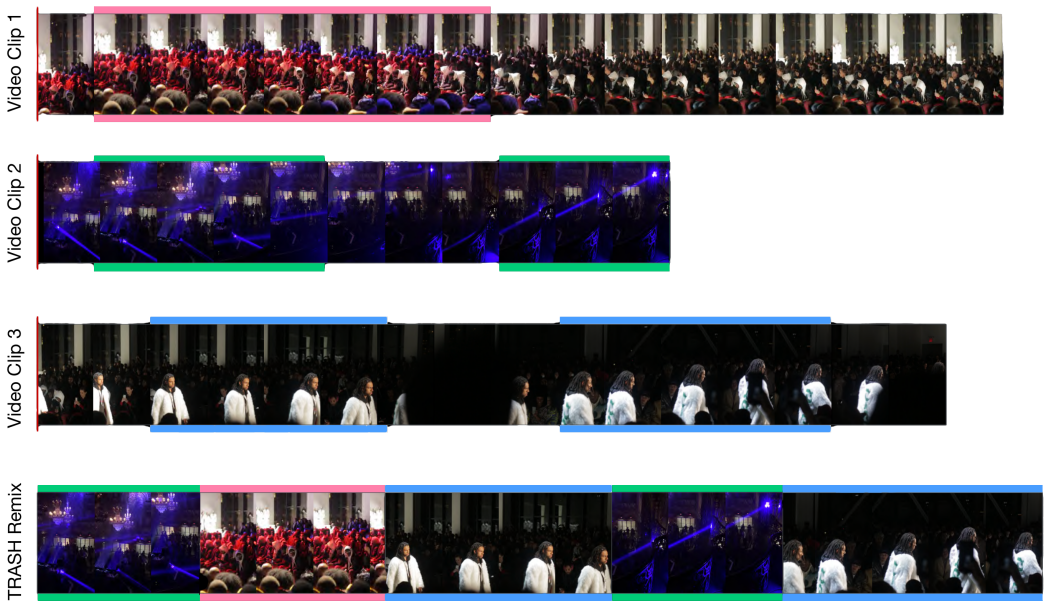

Figure 3: **Visualization of TRASH Editor Workflow.** The above figure shows three video clips shot at a fashion show by a test user of TRASH. Clip 1 shows a view of the audience (detected Wide Shot, Audience), clip 2 shows a view of the chandelier lighting the venue (detected Wide Shot), and clip 3 shows one of the models walking down the catwalk (detected Person Posing). The highlighted sections of each clip were automatically identified as interesting or important sections by the TRASH computational cinematography detectors and remixed to highlight these moments. The final remixed clip was arranged from the trimmed input clips.

BIOGRAPHIES

FRANCK NGAMKAN (SPEAKER)

Franck is a Research Scientist at the early-stage startup TRASH Inc. He recently completed a Master in Data Science at Columbia University and received a Master in Applied Mathematics from Ecole Polytechnique prior to that. His research focuses on audio signal understanding and narrative sequence generation.

GENEVIÈVE PATTERSON

Geneviève is Chief Scientist at TRASH. Previously, she was a Postdoctoral Researcher at Microsoft Research New England. Her work is about creating dialog between AI and people. Her interests include video understanding, computational cinematography, and human-in-the-loop AI systems. She received her PhD from Brown University in 2016 under the direction of James Hays.

REFERENCES

David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, 2018.

Vineet Gandhi, Remi Ronfard, and Michael Gleicher. Multi-clip video editing from a single viewpoint. In *Proceedings of the 11th European Conference on Visual Media Production*, pp. 9. ACM, 2014.

Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, 2018.

Alexandre Kaspar, Geneviève Patterson, Changil Kim, Yagiz Aksoy, Wojciech Matusik, and Mohamed Elgharib. Crowd-guided ensembles: How can we choreograph crowd workers for video segmentation? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 111. ACM, 2018.

Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Transactions on Graphics (TOG)*, 36(130), 2017.

Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

Billal Merabti, Marc Christie, and Kadi Bouatouch. A virtual director using hidden markov models. In *Computer Graphics Forum*, volume 35, pp. 51–67. Wiley Online Library, 2016.

Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, 2018.

Genevieve Patterson, Grant Van Horn, Serge J Belongie, Pietro Perona, and James Hays. Tropel: Crowdsourcing detectors with minimal training. In *HCOMP*, pp. 150–159, 2015.

Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. Quickcut: An interactive tool for editing narrated video. In *UIST*, pp. 497–507, 2016.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.