Turning an Urban Scene Video into a Cinemagraph Supplementary Material

Hang Yan
Washington University
St. Louis, USA
yanhang@wustl.edu

Yebin Liu
Tsinghua University
Beijing, China
liuyebin@mail.tsinghua.edu.cn

Yasutaka Furukawa
Washington University
St. Louis, USA
furukawa@wustl.edu

The supplementary document provides more details on the classification algorithm for randomly changing appearances.

1. Classifying randomly changing appearances

We use a binary random forest to identify segments of randomly changing appearance that lead to high quality animations. Table. 1 gives the full specification of our feature vector that encodes appearances, temporal, shape and position information in an 2D segment.

We have obtained the training data as follows. First, we have downloaded stationary video footages from YouTube and manually annotated 2D display regions at pixel levels. Second, we have run the same video segmentation algorithm and generate segments. Lastly, we label each segment as a positive (resp. negative) sample if more (resp. less) than 80% the segment overlaps with the annotated display pixels. We then train a random forest (100 decision trees with depth 10) by using 48 video clips for training and 12 video clips for validation. The training and validation accuracy are 98.0% and 93.0%, respectively. After obtaining the optimal hyperparameters, we merge the validation set into the training set and re-run the training.

At test time, we classify each segment by the trained random forest. Since image segments are obtained from the three levels of granularity, each pixel belongs to three different image segments. We treat a pixel to be a dynamic appearance segment, if at least one of the three segments is classified as a display. After finding the connected display components, we discard too small (less than 50 pixels) or too large (more than 30% of the frame) segments. Finally, we also discard segments that are mostly occluded in the other views and do not likely produce good animations. More precisely, we discard a segment if more than half the pixels are occluded in more than half the neighboring views based on the visibility test conducted in the video warping process.

References

[1] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008. 2

Category	Feature	description	Dim
Appearance	RGB mean	Mean RGB values.	3
	RGB variance	Variances of RGB values.	3
	LAB histogram	Histogram in LAB color space with 8 bins for each channel.	24
Gradient	BoW	Bag-of-words descriptors constructed by K-means clustering on HoG3D descriptor [1] extracted from regular grid points.	100
Shape	Area	Ratio of the 2D area against the area of the entire frame.	1
	Convexity	Ratio of the 2D area against the area of its convex hull.	1
	Rectangleness	Ratio of the 2D area against the area of the minimum bounding box.	1
	Aspect ratio	The aspect ratio of the minimum bounding box.	1
	Number of edges	The number of edges of an approximated 2D shape. The approximation error is set to 1% of the smaller dimension of the frame.	1
Position	Centroid	The position of the centroid of the segment.	2
	Bounding box	Minimum/maximum x/y position, normalized by width and height.	4

Table 1: The feature vector used for random forest.