https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-283

Practical Automatic Thumbnail Generation for Short Videos

Bin Shen, Nikil Pancha, Andrew Zhai, Charles Rosenberg
Pinterest Inc.

Abstract

With the availability of fast internet and convenient imaging devices such as smart phones, videos are becoming increasingly popular and important content on social media platforms recently. They are widely adopted for various purposes including, but not limited to, advertisement, education and entertainment. One important problem in understanding videos is thumbnail generation, which involves selecting one or a few images, typically frames, which are representative of the given video. These thumbnails can then be used not only as a summary display for videos, but also for representing them in downstream content models. Thus, thumbnail selection plays an important role in a user's experience when exploring and consuming videos. Due to the large scale of data, automatic thumbnail generation methods are desired since it is impossible to manually select thumbnails for all videos. In this paper, we propose a practical thumbnail generation method. Our method is designed in a way that will select representative and high-quality frames as thumbnails. Specifically, to capture semantic information of video frames, we leverage the embeddings of video frames generated by a state of the art convolutional neural network pretrained in a supervised manner on external image data, using them to find representative frames in a semantic space. To efficiently evaluate the quality of each frame, we train a linear model on top of the embeddings to predict quality instead of computing it from raw pixels. We conduct experiments on real videos and show the proposed algorithm is able to generate relevant and engaging thumbnails.

Introduction

With the rapid development of information technology such as high-speed networks and widely available imaging devices, there is an increasing amount of multimedia data on the web. Videos are becoming much more important content on social media platforms. For example, hundreds of hours of video are uploaded to YouTube every single minute [1], and the number is still increasing. Videos play critical roles in many industries, such as entertainment, education and advertisement, since they offer much richer information than static images.

To better understand and utilize videos, a natural approach is to find one or a few representative images, which are typically frames of the same video, to represent each of the videos. These representative images are called thumbnails and can be used for multiple purposes. The primary purpose of these is often for visualizing a video in situations where playing it might be expensive or inappropriate, but, beyond the visible, they can be used as a signal in downstream applications. For example, they can provide a succinct representation of a video's content, allowing machine learning models to process only a few images, rather than thousands of frames, and they can allow for models trained to handle images to be directly applied to videos

Although there are many applications of video thumbnails as mentioned above, not all videos are created with specified thumbnails. As corpora of videos become larger, it becomes prohibitively expensive to manually pick thumbnails for all of them. Thus, we propose an automatic thumbnail generation method to pick one or multiple frames as thumbnail(s) for each video.

The proposed thumbnail generation method is designed to pick high-quality and representative thumbnails for each video. It addresses the following practical issues: (1) How can a model analyze the video in semantic space by transferring external knowledge to video domain? (2) How can we measure image quality efficiently without directly using raw pixels? To evaluate the proposed algorithm, we conduct experiments on videos and visualize thumbnails, which will show that the proposed algorithm is able to generate pleasing thumbnails. Moreover, for the application of multiple thumbnail generation, our method is able to generate a diverse thumbnail set for each video. And, importantly, We also conduct experiments in production systems that use the generated thumbnails, and show they can benefit from the thumbnails in at least two applications: (1) as a method to preview videos; (2) as a signal in downstream machine learning models.

Related Works

Here we review a few related works on thumbnail selection and a related task video summarization.

A set of prior work proposes to select video thumbnails by designing criteria that explore low-level features, such as color and motion [2, 3]. [4] segments videos according to camera motion and then uses rules to select a key frame for each segment, followed by a semantic relevance ranking module. Also, there are other works that use high level features, such as face detection results, which incorporate semantic information into thumbnail selection [5]. [6] considers not only image features, but also textual information. It treats thumbnail selection as a ranking problem, and ranks the frames according to their relevance to the video theme with the help of keywords which are used retrieve similar images for theme analysis. [7] also uses the textual keyword/tag information available for the video to identify relevant frames with a topic model and relevance filtering. Similarly, multiple instance learning is used in [8] to localize textual tags. Saliency is introduced for video summarization in [9], which proposes to predict the importance score for frames by using a simple linear model. However, the semantic information is generally ignored in this approach. [10] introduces a deep semantic embedding model for thumbnail selection. In particular it maps the text and image into the same embedding space, allowing the similarity between text and image can be computed.

To compare our proposed work to the existing works, the following aspects should be noted:

- Our method also uses deep semantic embeddings. However, we use a more powerful neural network, which is trained using multi-task learning causing the network to learn a common representation for different tasks [11, 12]. Moreover, we use binary embeddings due to the potential efficiency in storage and computation.
- Similar to [6], which relies on external images for theme analysis, our method also relies on external data. The difference is that in our case, the external data is used to train the convolutional neural network, which computes the semantic embeddings for video frames. Thus, the information of external data is incorporated into the model.
- Our method does scene analysis, which is similar to theme analysis, but we do not depend on keywords or other forms of text, which may not be available or informative for all videos. Instead, we directly cluster the frame embeddings to learn the scenes.

Proposed Thumbnail Generation Method

To automatically generate thumbnails for videos, the following aspects are considered: (1) frame representativeness; (2) frame quality; (3) diversity if more than one thumbnail is generated. Not that we use the terms frame and image interchangeably except in cases where they need to be differentiated.

To model the representativeness of a frame, it is natural to consider how many frames in total in a video are similar to it. Thus, we propose to cluster all candidate frames in semantic space, and each cluster could be viewed as a scene. The size of the cluster matters. The frames belonging to a larger cluster are more representative of the video than the ones belonging to a smaller cluster. However, clustering frames in semantic space may be difficult. Thus we propose to use external resources, which help train a convolutional neural network (CNN) to compute semantic binary embeddings for candidate frames. The binary embeddings not only pave the way for further analysis in semantic space, but also enjoy potential efficiency in storage and computation. To explicitly model frame quality, we propose to train a linear model, the details of which will be discussed in a following section. Finally, the MMR method [13] is introduced to enforce diversity if more than one thumbnails are selected. Figure 1 illustrates the proposed method, which has the key components: frame feature extraction, image quality model, scene analysis and representative image ranking.

Semantic Embedding Extraction

Unlike many traditional methods, which focus on low-level features, we believe that semantics of frames could be useful in thumbnail selection. For example, with the semantic understanding, the algorithm will be able to understand that different frames may be about the same semantic things.

There are multiple ways to derive semantics for frames. One possible approach is to use machine learning algorithms to annotate each frame with semantic tags. However, this approach may be sensitive to the annotation algorithm. Another alternative is to collect text tags for videos. However, this textual information may not be available for many videos. Thus, we decide to use semantic embeddings, which are extracted for all frames for semantic analysis.

It can be difficult to train a model, such as a convolutional neural network, from video frames to generate semantic embeddings due to the lack of labeled data. However, in real world applications, there is a large volume of image data with labels from other tasks. Many works on multi-task learning explore a common representation of different tasks [11, 12], thus we believe embedding models trained on these non-video data could then applied to video frames.

More specifically, we use a convolutional neural network pretrained by the multi-task learning based approach [12], which uses SE-ResNeXt101 [14] as a base model, with the goal that the embedding learned in this way could be transferred to video frame domain. With the help of this pre-trained neural network, each frame is represented as a 2048 dimensional binary feature vector. Mathematically, for each frame $\mathbf{x}$, the network computes a semantic embedding $\mathbf{e}(\mathbf{x}) \in \{0,1\}^{2048}$.

Image Quality Model

Image quality is important for video thumbnails, especially for visualization purposes. A general definition of image quality is "the weighted combination of all of the visually significant attributes of an image" [15]. The exact definition of image quality depends on the specific application. In our case, the quality of an image is defined so as to solve the practical problem of blank thumbnails. We observe that in real systems many default thumbnails are blank or almost blank images, which is undesirable since they are typically not attractive to users and do not provide much information for potential downstream machine learning algorithms. A reason for having this kind of blank images as thumbnails is that in some systems the first frames are treated as the default thumbnail while the first frames of many videos are blank. Sample blank thumbnails are shown in Figure 2.

An image with width W, height H, and number of channels C, can be represented as a d dimensional vector $\mathbf{x}$ by vectorizing the pixel values. Mathematically, $\mathbf{x} \in R^d$, where $d = W \times H \times C$. Let $\mathbf{x}(I(w,h,i))$ denote the value of the i_{th} channel of the pixel at position (w,h), where $I(.)$ is the mapping function that maps (w,h,i) to the corresponding index of $\mathbf{x}$.

It is possible to compute the image blankness $b(\mathbf{x})$ purely based on pixels using low-level image processing techniques. The blankness could be defined based on the $L2$ norm of the image gradient, which could be noted as:

$$b(\mathbf{x}) = 1 - \frac{\|concat(\frac{d\mathbf{x}}{dw}, \frac{d\mathbf{x}}{dh})\|}{Z}, \tag{1}$$

The gradient information of the image is encoded in $\frac{d\mathbf{x}}{dw}$ and $\frac{d\mathbf{x}}{dh}$), and $concat(\frac{d\mathbf{x}}{dw}, \frac{d\mathbf{x}}{dh})$ is the resulting vector formed by concatenating $\frac{d\mathbf{x}}{dw}$ and $\frac{d\mathbf{x}}{dh}$. In the equation above, Z is a normalization constant to ensure the $b(\mathbf{x})$ is in the range of $[0,1]$. In our implementation, 8-bit value is used to represent each pixel value $\mathbf{x}(I(w,h,i))$. As a result, the derivatives $\frac{d\mathbf{x}(I(w,h,i))}{dw}$ and $\frac{d\mathbf{x}(I(w,h,i))}{dh}$ are also bounded. Z could be set as the maximum possible value of $\|concat(\frac{d\mathbf{x}}{dw}, \frac{d\mathbf{x}}{dh})\|$. For example, Z is set to $2 \times 256 \times C \times W \times H$ for the 8-bit case.

However, due to the large number of pixels that they have to operate on, the computation will be expensive and slow.

To speed up the computation of the quality, we propose to predict the blankness of image $\mathbf{x}$, $b(\mathbf{x})$, from its semantic embedding, $\mathbf{e}(\mathbf{x}) \in \{0,1\}^{2048}$.

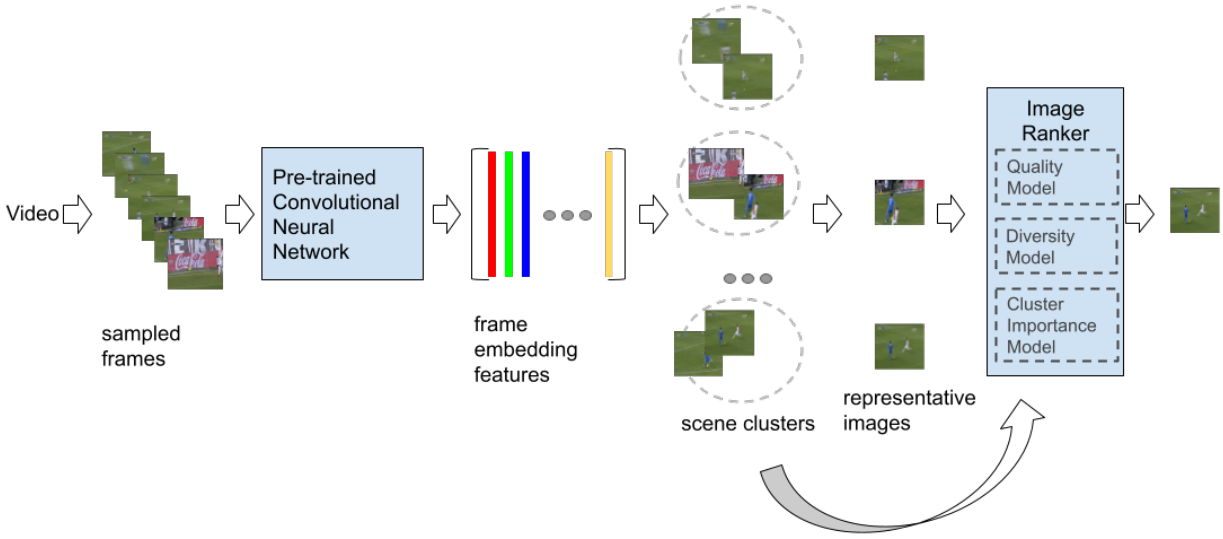


Figure 1: Algorithm overview

More specifically, we collect training data, a set of video thumbnails, and then compute the blankness of these images using Equation 1. The thumbnails with high blankness scores are labeled as *blank*, and the ones with low scores are labeled as $non-blank$. A linear logistic regression model, which takes embedding $\mathbf{e}(\mathbf{x})$ as input, is trained from the data to predict the label of the thumbnail. By doing this, the algorithm will execute a fast computation of image quality without touching pixels, assuming the embedding is given, and the computation cost is reduced from $O(W \times H)$ to $O(1)$ compared with the naive approach which computes blankness from pixels. The reason why we assume the embeddings are given and do not count the cost is that the other component of the algorithm, the clustering part, already requires the embeddings of all the frames. Thus, the quality prediction module gets the embeddings for free.

Figure 2: Sample blank thumbnails

Scene Analysis

Scene Discovery via Clustering The candidate frames are clustered to k clusters, where k could be empirically set as $\min(\max(\frac{N_{frames}}{T}, N_{thumbnails}+c), N_{frames})$, where N_{frames} is the number of input frames, $N_{thumbnails}$ is the desired number of thumbnails, and T and c are parameters. T controls the average size of clusters when there are many candidate frames, while c ensures there are enough clusters when there are few frames. The clustering is conducted using the semantic embedding, so the frames in the same cluster are semantically similar to each other.

Representative Image Selection Within Cluster Our method picks one representative frame for each cluster. More specifically, the image that is closest to the clustering center is chosen as the representative image for that cluster. Considering that the clustering center probably does not have a binary representation, the distance between a cluster center and an image is measured by Euclidean distance. For a cluster i, the representative image is denoted as $representative\_image(i)$. Also, we define its importance as:

$$\text{importance}(i) = w(size(i))\text{quality}(representative\_image(i)), \quad (2)$$

where $size(i)$ is the size of cluster i, $w(.)$ is a positive non-decreasing monotonic function. In our implementation, $w(.) = \sqrt{.}$, $\text{quality}(.) = 1 - b(.)$, where $b(.)$ could be approximated by logistic regression as discussion above.

Intercluster Similarity When more than one thumbnail is desired, the diversity of the final selected thumbnails should be considered. Considering the frames are represented as binary embeddings, the distance between two frames i and j can be easily measured by the hamming distance, which is denoted $d(i,j)$. Let $\mathbb{F}$ denote the set of representative images for all clusters. Since there is one representative image per cluster, $|\mathbb{F}| = k$.

To count the different variation of different videos, we compute the normalized distance $\tilde{d}(i,j)$ between any pair of representative images i and j in $\mathbb{F}$ as below:

$$\tilde{d}(i,j) = \frac{d(i,j)}{\max_{i',j' \in \mathbb{F}} d(i',j')}. \quad (3)$$

The similarity of between them is defined as $sim(i,j) = 1 - \tilde{d}(i,j)$.

Representative Image Ranking

Given k representative images in $\mathbb{F}$, it is necessary to sort them considering both the importance and diversity when more than one thumbnails are desired. To achieve this goal, Maximum Marginal Relvance (MMR) [13] is introduced as the ranking method. Let $\mathbb{S}$ denote the set of selected thumbnails. At the beginning, $\mathbb{S}$ is initialized to an empty set $\emptyset$. At each step, MMR greedily picks the best representative image from $\mathbb{F}$ and adds it to $\mathbb{S}$. More specially, MMR picks the first image by maximizing $\max_{i\in\mathbb{F}} \text{importance}(i)$. When $\mathbb{S}$ is not empty, MMR picks the image by maximizing the following objective function:

$$i' = \arg\max_{i\in\mathbb{F}\setminus\mathbb{S}} \lambda \,\text{importance}(i) - (1-\lambda)\max_{j\in\mathbb{S}} sim(i,j) \qquad (4)$$

where $\lambda \in [0,1]$ is parameter controlling the trade-off. It is easy to see that if only one image is required then the most important image will be picked. If more than one image is required, then the images selected later are required to be important and different enough from the previously selected ones. By doing this, representative images from near duplicate clusters are suppressed when the algorithm searches for the most important representative images.

Experiments

First, we evaluate a component of the proposed method, the quality prediction model. Then, the proposed method is evaluated for the thumbnail generation task.

Image Quality Prediction

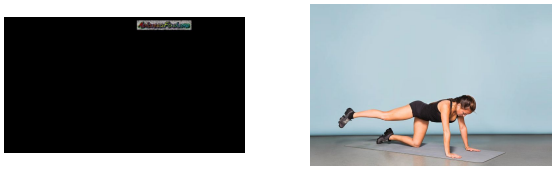
For the experiment of image quality prediction, 491974 videos are collected as training data, each of which has a default thumbnail, resulting in a total of 491974 thumbnails. The goal is to train a quality prediction model from this set of thumbnails.

However, it is still expensive to manually annotate the quality of all the thumbnails. Thus, to annotate the thumbnails, as mentioned above, we compute the blankness of these images with Equation 1. Then all the thumbnails are sorted according to the computed blankness. We label the top 20000 thumbnails as *blank* and the bottom 20000 thumbnails as $non-blank$. It is difficult to label the thumbnails in the middle, so we only label the thumbnails that we are confident about, which are the top and bottom ones.

A linear logistic regression model is trained on randomly selected 90% of the data and evaluated on the rest. The accuracy is over 99%. Figure 3 shows the blankness scores predicted by the model for one positive example and one negative example. As we can see, it correctly predicts the blank image even though there is some text, and correctly predicts the non-blank image even though most of the areas are smooth.

Thumbnail Visualization

The proposed approach is also experimented with in our real system. We compare the proposed the method with a baseline approach, which relies on the user's selection. When a user uploads a video, the system takes the first frame as the default thumbnail and allows the user to make change to the selection. Thumbnails generated by this baseline approach and the proposed approach are compared in Figure 4. As we can see, the thumbnails generated by our proposed approach generally look more pleasing than



(a) 0.999999990789 (b) 0.00202098354575

Figure 3: Blankness prediction results of two example images

the ones from the baseline approach. Also, as expected, the ones generated by our method contain richer information, while there are some nearly blank thumbnails displayed by the baseline approach.

Moreover, one merit of our algorithm is that it can select multiple thumbnails. To demonstrate its effectiveness, three thumbnails are selected and then visualized for sample videos as shown in Figure 5. The thumbnails of the recipe video are able to show the different phases of the food preparation; the ones of yoga video show three different poses; and the ones of dress video show different dresses. The selected multiple thumbnails demonstrate sufficient diversity, which allows them to provide complementary information for each other.

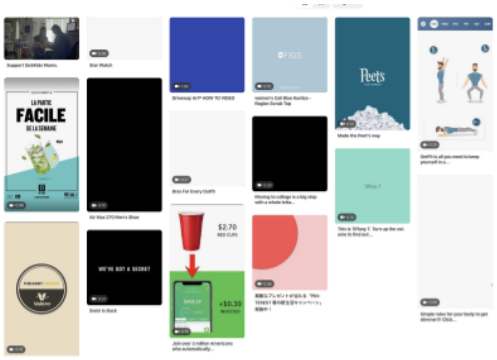
Quantitative Comparison

We are primarily interested in two main use cases for thumbnails: (1), as a way to display videos; (2), as signals for videos. It is difficult to qualitatively evaluate how well the generated thumbnails are without conducting expensive user studies. However, the thumbnails can be more easily quantitatively evaluated in these two use cases.

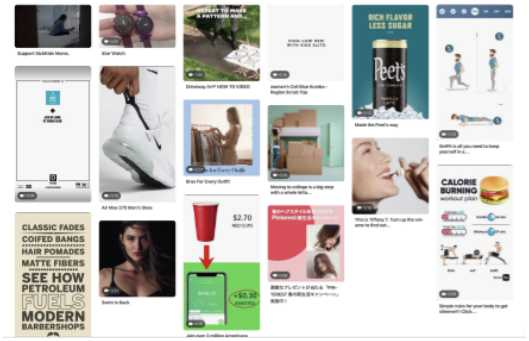
In our video system, we would like to respect the user's selection of thumbnails. Thus, we try to not replace user's selection. However, when the thumbnail for a video in system is blank, which probably means that the user did not select it., we use our algorithm to select it.

In our experiment, we use the same set of 491974 videos as the data set for evaluation. Each of the videos comes with a thumbnail generated by the baseline method. We sort thumbnails according to blankness score computed in Equation 1. The top 20000 videos according to blankess score are selected for comparison. We use our proposed method to generate thumbnails for this selected set of videos, and compare them with the thumbnails in system, which are generated by the baseline approach.

Thumbnails are displayed for the videos. Two types of thumbnails are compared in this use case. More specifically, two randomly selected groups of users are selected. One group of users are shown thumbnails generated by the baseline approach while the other group of users see the thumbnails by our method. If one user does not like the video shown, she/he has the option to hide the video. Thus, the hide rate can be viewed as a metric to measure how well the thumbnail is. If a user likes the video shown, he/she has the option to click on the video, expanding it to take up a larger portion of the screen. Thus the close up rate could be treated as another metric. In our experiment, we see that switching from the default thumbnails in the existing system to the thumbnails by our methods, the hide rate of videos is decreased by 4%. Also, the closeup rate goes up by 1.7% relatively.



(a) Thumbnails by the baseline approach



(b) Thumbnails by the proposed approach

Figure 4: Thumbnail visualization

(a) Recipe thumbnail 1

(b) Recipe thumbnail 2

(c) Recipe thumbnail 3

(d) Yoga thumbnail 1

(e) Yoga thumbnail 2

(f) Yoga thumbnail 3

(g) Dress thumbnail 1

(h) Dress thumbnail 2

(i) Dress thumbnail 3

Figure 5: Multi-thumbnail selection of sample videos

Video features can be computed by combining the thumbnail generation algorithm with downstream machine learning algorithms. For example, with Graph Convolutional Neural Network [16], the embedding of the selected thumbnail for the video could be combined with other features, such as text and visual information of its neighbors. Thus, we compute video embeddings using the method in [16], and generate two versions of the embeddings: the first version uses the visual embeddings of the thumbnails generated by the baseline approach, and the second version uses the embeddings of the thumbnails generated by our proposed method. The videos are indexed by these two versions of visual embeddings, and fed to the serving system. Then we utilize the closeup and clickthrough rates as metrics. We found both of these metrics to be statistically significantly improved; both video closeup rate and clickthrough rate are increased by 2% relatively by using our method.

Conclusion and Discussion

We propose a practical automatic thumbnail generation method. First, semantic embeddings trained from external multi-task data are used to power the thumbnail generation. Second, a fast image quality model is trained so that we can quickly evaluate a thumbnail's quality without working at the raw pixel level. And last but not least, when an application requires more than one thumbnail to be selected, our method is able to trade off quality and diversity. Experimental results on real videos show that the generated thumbnails not only qualitatively look pleasing, but also help improve the browsing and exploration experiences for users.

For future research, it would be interesting to use more supervision information to improve the thumbnail selection. Instead of using expensive strong supervision information, we may consider using some format of weak supervision information. For example, a user may click a video immediately after she or he browses an image. Thus, we may assume with high probability this image is similar to the thumbnails of the video. If this assumption is valid, then we may collect image and video pairs from log information, and train a model using this weak supervision to learn better video thumbnails.

References

[1] https://blog.youtube/press

[2] Yihong Gong, Xin Liu, Generating Optimal Video Summaries, IEEE International Conference on Multimedia and Expo 2000.

[3] Xian-Sheng Hua, Shipeng Li, Hong-Jiang zhang, Video Booklet, IEEE International Conference on Multimedia and Expo 2005.

[4] Jiebo Luo, Christophe Papin, Kathleen Costello, Towards Extracting Semantically Meaningful Key Frames From Personal Video Clips: From Humans to Computers, IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 2, pp. 289-301, Feb. 2009.
[5] Frederic Dirfaux, Key Frame Selection To Represent A Video, International Conference on Image Processing 2000.
[6] Yuli Gao, Tong Zhang and Jun Xiao, Thematic Video Thumbnail Selection, IEEE International Conference on Image Processing 2009.
[7] Haojie Li, Lei Yi, Bin Liu, Yi Wang . Localizing relevant frames in web videos using topic model and relevance filtering. Machine Vision and Applications 25, 1661–1670, 2014.
[8] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, Tat-Seng Chua, Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification, in IEEE Transactions on Multimedia, vol. 14, no. 4, pp. 975-985, Aug. 2012.
[9] Zheng Lu, Kristen Grauman, Story-Driven Summarization for Egocentric Video, IEEE Conference on Computer Vision and Pattern Recognition 2013.
[10] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, Jiebo Luo, Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection, IEEE Conference on Computer Vision and Pattern Recognition 2015.
[11] Shibin Parameswaran, Kilian Q. Weinberger, Large Margin Multi-task Metric Learning. In Conference on Neural Information Processing Systems, pages 1867–1875, 2010.
[12] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, Charles Rosenberg, Learning a Unified Embedding for Visual Search at Pinterest, ACM SIGKDD International Conference on Knowledge Discovery and Data 2019.
[13] Jaime Carbonell, Jade Goldstein, The Use of MMR, Diversity-Based Reranking For Reordering Documents And Producing Summaries, the 21st annual international ACM SIGIR conference on Research and development in information retrieval 1998.
[14] Jie Hu, Li Shen, Gang Sun, Squeeze-and-Excitation Networks, IEEE Conference on Computer Vision and Pattern Recognition 2018.
[15] Norman Burningham, Zygmunt Pizlo; Jan P. Allebach, Image Quality Metrics. In Hornak, Joseph P. (ed.). Encyclopedia of imaging science and technology, New York: Wiley, 2002.
[16] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, Jure Leskovec, Graph Convolutional Neural Networks for Web-Scale Recommender Systems, ACM SIGKDD International Conference on Knowledge Discovery and Data 2018.

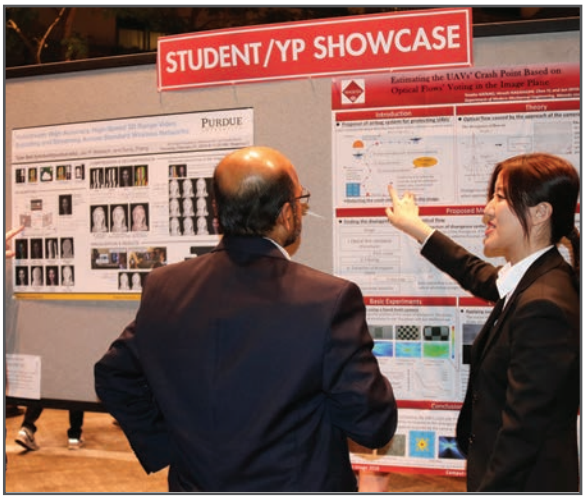
JOIN US AT THE NEXT EI!

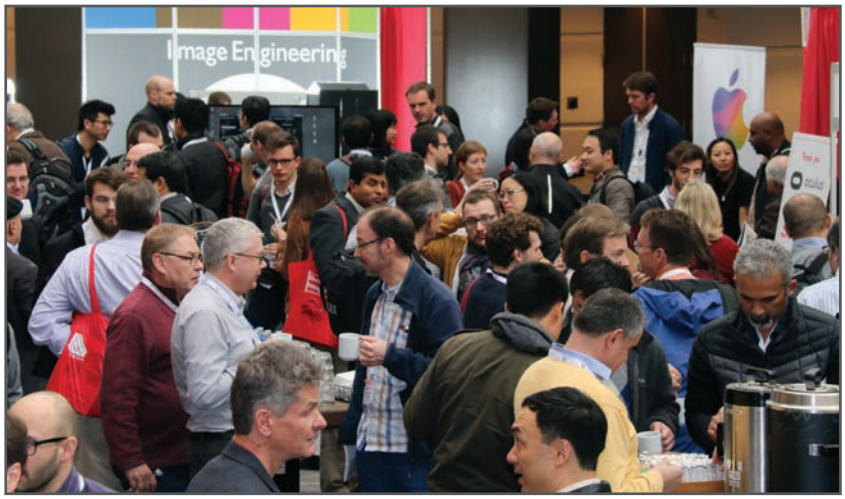
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org