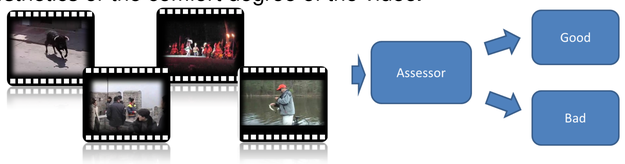
Video Aesthetic Quality Assessment by Combining Semantically independent and dependent features

Chun-Yu Yang[1], Hsin-Ho Yeh[1] and Chu-Song Chen[1,2]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan.
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.

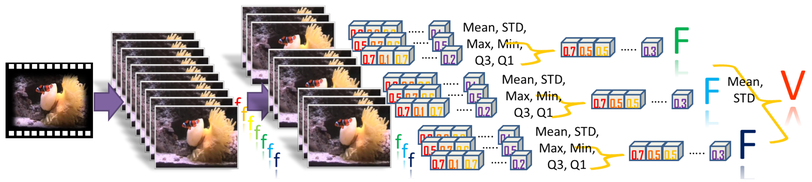
Background

- Video Quality Assessment
  - High quality videos have been the pursuit of human beings. In an HD-video-widely-spread era, to judge the quality of a video would mostly depend on the aesthetics or the comfort degree of the video.



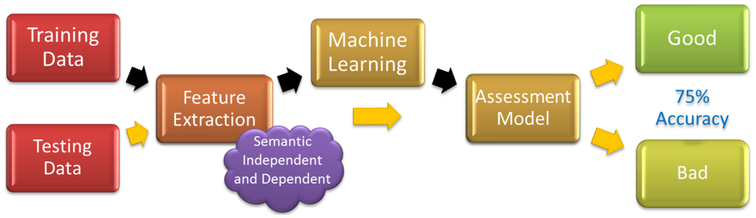
What Others Have Done

- Photo-based Assessor
  - Aesthetic quality assessment had been studied thoroughly in photos. They focus on extracting aesthetic features to represent the artistic feeling a person perceive toward photos.
- Temporal Property Combined Assessor [Moorthy et al. in ECCV10]
  - Moorthy et al. propose a hierarchical pooling method by combining photo-based aesthetic features and temporal property to model the aesthetic quality of videos.



What We Have Done

- Motion related aesthetic features imported
  - Previous work considers only the motion in finding handshaking, we further introduce some important features: motion space and motion direction entropy.
- Analysis in aesthetic features by introducing semantic property
  - Some features have the property that aesthetic criteria measured from them vary with the video content. This we call them semantic dependent features.



- Combining the newly explored features with the existing features to learn an assessment model by SVM, we achieve the state-of-the-art performance which has 75% accuracy.

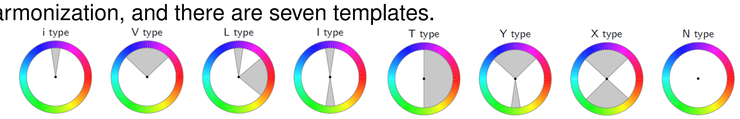
Semantic-Independent Feature

- Motion Space:
  - The space in front of the moving direction of the object should be reserved for the better aesthetic quality of the video, since it gives the audience more imagination about the subject. The feature of motion space is represented as $\boldsymbol{f = v \cdot d}$.

- Hand-shaking:
  - We simulate the shaking effect as follows, the motion indicator $\boldsymbol{I_j}$ in frame $\boldsymbol{j}$ is defined as an unit step function $\boldsymbol{I_j = u(mx_j)}$ where $\boldsymbol{mx_j}$ represents the border motion vector in frame $\boldsymbol{j}$ along the horizontal direction. The magnitude of motion indicates the degree of unstableness; thus horizontal unstableness feature $\boldsymbol{f}$ is defined as:

$$\boldsymbol{f = (I_j \oplus I_{j-1}) \times (|mx_j| + |mx_{j-1}|)}.$$

- Colour Harmonic [Cohen-Or et al in ACM Trans.06]:
  - The human visual perception of aesthetics is strongly related to colour harmonization, and there are seven templates.



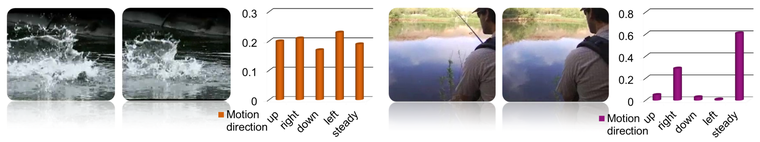
$$\boldsymbol{f_c = |h *' Tp - Tp_c *' Tp|^2, \forall c \in [1, 2, ..., 7]}$$

, where $\boldsymbol{Tp = [Tp_1, Tp_2, ..., Tp_7]^T}$ and $*'$ means to convolute and choose the maximum value of the resulted histogram.

- Composition:
  - The composition of a frame is also of great importance in an aesthetic view, in this work we adopt the rule of third, clarity contrast, and shape convexity.

Semantic-Dependent Feature

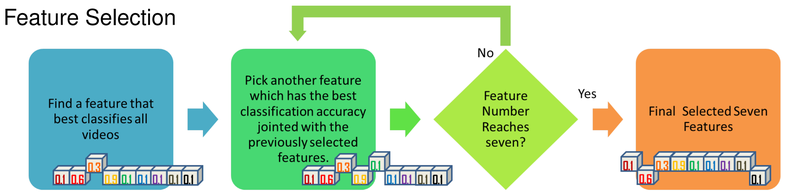
- Motion Direction Entropy:
  - Entropy is commonly treated as the amount of uncertainty, whereas the concept of the entropy is used to quantify the uncertainty of the motion direction of every pixel in each frame. $\boldsymbol{f = \sum_{b=1}^{5} -p_b \ln(p_b)}$, where $\boldsymbol{p_b = \frac{bin_b}{\sum_{k=1}^{5} bin_k}}$



- Colour Saturation and Value:
  - In HSV space, there are saturation and value to be considered; thus we compute the average saturation and value for each frame as additional colour features. And we add two more colour features by averaging the centre block.
- Lightness:
  - we define one lightness feature as the lightness ratio of subject and background without subject. The other feature is described as the lightness ratio of subject and whole frame.

Experiments Definition

- Dataset Introduction:
  - This dataset consists of 160 videos with 15 seconds short-segment, and each video was rated by two authors on a 5-point scale[Moorthy et al. in ECCV10].
- Evaluation
  - Using 5-fold cross-validation and repeating it 200 times to obtain assessment accuracy.
  - Feature Selection



Experimental Result

- Assessment performance using semantically dependent and independent features separately

semantic-dependent	semantic-independent
69±2.2%	**74 ± 1.5%**

- Accuracy obtained by combining the two types of features

I. / D.	Motion	Colour	Lightness	Comp.
Motion	72±1.3	63±2.1	**60 ± 2.3**	**66 ± 1.5**
Colour	**74 ± 1.4**	64±2.5	56±2.2	60±2.2
Lightness	71±1.7	**70 ± 2.1**	59±2.4	65±2.4
Comp.	72±1.3	63±2.7	N/A	65±2.5

- Finally, we use the total 22 features for assessment. To show the improvement using our method, two works are carried out for comparison: Moorthy's method and Luo's method[Luo et al in ECCV08]. The assessment performances are **75 ± 1.1%**, 73±2.0% and 54% in our method, Moorthy's method and Luo's method, respectively

Conclusion and Future Work

- Currently, our method has a 75% accuracy in distinguishing good or bad videos and we wish to accomplish a scoring system which allows us to carry out further application in near future for example an automatic aesthetic detector/tutor while videotaping.