EURECOM
Sophia Antipolis

VERT: Automatic Evaluation of Video Summaries

Yingbo Li, Bernard Merialdo

EURECOM, France

AIEMPro10 Workshop, Florence, 29 Oct 2010

Introduction

- Review of video summarization
- Evaluation of video summaries
- BLEU and ROUGE
- VERT principles
- Experiments
- Conclusion

EURECOM - BP 193
F-06904 Sophia Antipolis cedex

Video Summarization

- **Overload of Multimedia information, specially videos**
  - Lots of TV channels
  - Lots of recording devices
- **Summarization is a useful tool:**
  - Quickly grasp the main content
  - Decide to watch entire video or not
  - Allows to quickly compare several videos
  - Sometimes find relevant information
- **Major issue in summarization:**

Video Summarization is difficult

- **Efficient selection requires:**
  - Analysis
  - Modeling
  - “Understanding”
  - Evaluation of importance

Video Summarization is easy

- **Lots of possible approaches for selection**
  - From random choice
  - To numerical optimization
- **How to prove that a summary is good (or bad)?**
- **A major problem is Evaluation**

Video Summary Evaluation

- **Many proposals, two basic approaches:**
- **Objective metrics (quantitative)**
  - *SVD over feature frame matrix [Gong 2000]*
  - *Shot Reconstruction Degree [Liu 2004]*
  - *Shot importance [Uchihashi 1999]*
- **User studies (qualitative)**
  - *Keyframe Counting [Dufaux 2000]*
  - *User satisfaction [Ngo 2003]*
  - *Content identification [Smith 1998, Lu 2004]*

Video Summary Evaluation

- **Problem with current approaches:**
  - Maximize objective metrics
    - Performance does not always relate easily to a task
    - Result is difficult to interpret
  - Evaluate with real users on real task
    - Very expensive, difficult to set up
    - Difficult to optimize summaries automatically
- **Fundamental difficulty:**
  - There is no ground truth
  - But people are able to judge if one proposal is better or worse than another

BLEU and ROUGE

- **BLEU (Bilingual Evaluation Understudy)**
  - A similar situation is encountered in language translation
  - Proposal: BLEU measure (IBM 2002)
  - Idea: measure the similarities between a candidate translation and a set of reference translations
    - Compare n-gram counts
    - Precision-based measure

$$BLEU_n = \frac{\sum_{C \in \{CandidateSentences\}} \sum_{gram_n \in C} Count_{clip}(gram_n)}{\sum_{C \in \{CandidateSentences\}} \sum_{gram_n \in C} Count(gram_n)}$$

  - High correlation with human judgment
  - Scoring metric used in the NIST translation benchmarks

BLEU and ROUGE

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**
  - Text summarization evaluation metric (Lin 2003)
  - Counts the number of overlapping units between the candidate summary and several man-made ground truth summaries

$$ROUGE-N = \frac{\sum_{S \in \{References\}} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S \in \{References\}} \sum_{gram_n \in S} count(gram_n)}$$

  - Recall oriented measure
  - Several variants:
    - ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S

VERT

- **VERT (Video Evaluation by Relevant Threshold)**
  - Transpose BLEU and ROUGE ideas to evaluation of video summarization
  - Issues:
    - Precision or Recall ?
    - How to define $gram_n$?
    - Which values of n ?
    - How to validate VERT ?
  - Video summary = selection of instants

    ~ selection of ordered keyframes

    $S = f_1 f_2 \ldots f_n$

n-gram word order ~keyframe rank (decreasing importance)

VERT-P

- **Inspired by BLEU, precision-based**
  - Keyframes are assigned a weight based on position in the selection
  - In reference summaries (human selected lists)
    - keyframe i in position y_i of reference x: $W_S(x,y_i)$
    - $T_i = \max_x W_S(x,y_i)$
  - In candidate summary (computer selected list)
    - keyframe i : $W_C(i)$
  - VERT-P:

$$VERT-P = \frac{\sum_{i=1}^{m} \min[W_C(i), T_i]}{\sum_{i=1}^{m} W_C(i)}$$

  - Maximal value when candidate keyframes all have a rank less or equal to their best rank in references

VERT-R

- **Inspired by ROUGE, recall based**
  - computes the weight percentage of reference $gram_n$ occurring also in the candidate summary

$$VERT-R_N(C) = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} W_C(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} W_S(gram_n)}$$

  - Variants:
    - N=1:

$$VERT-R_1(C) = \frac{\sum_{S \in R} \sum_{f \in S} W_C(f)}{\sum_{S \in R} \sum_{f \in S} W_S(f)}$$

    - N=2:

$$VERT-R_2(C) = \frac{\sum_{S \in R} \sum_{(f,g) \in S} W_C(f,g)}{\sum_{S \in R} \sum_{(f,g) \in S} W_S(f,g)}$$

VERT-R_{2S}

$$W_S(f,g) = \frac{w_S(f) + w_S(g)}{2}$$

VERT-R_{2D}

$$W_S(f,g) = |w_S(f) - w_S(g)|$$

Experiments

- **Videos related to news articles**
  - Obtained from Wikio web site
  - 2 groups of 6 videos each
  - 10 keyframes max per video

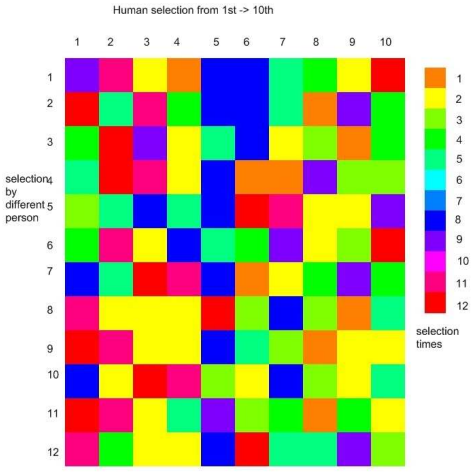
DATI

YSL

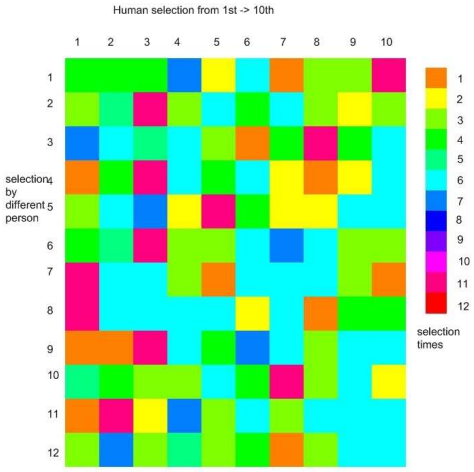
- **Reference summaries:**
  - 12 users: ordered selection of 10 keyframes

Experiments

- **User selection (12 users)**



12 Human selections of DATI



12 Human selections of YSL

Evaluating the evaluation method

- **Goal: compare VERT score with human judgement**
  1. Select 7 candidates summaries:
     - 2 random summaries
     - 1 summary constructed by K-Means
     - 2 summaries constructed by Video-MMR
     - best and worst human summaries
  2. Create 21 pairs:

Summary Pair: One row = one summary.

Evaluating the evaluation method

3. Request users to perform Human Pair Selection (HPS): select the best one for each summary pair
4. Use VERT to perform VERT Pair Selection (VPS) for each summary pair
5. Compare HPS and VPS:
   - Accuracy percentage λ: percentage of correct choices made by VPS compared with HPS

$$\lambda = \frac{1}{H}\sum_{h=1}^{H}\left[\frac{1}{21}\sum_{i=1}^{21}\frac{C_{VERT}(i)\cdot C_{H_h}(i)+1}{2}\right]$$

$$\text{where } C_X(i) = \begin{cases} -1 & \text{if the first summary is selected} \\ +1 & \text{if the second summary is selected} \end{cases}$$

   - Spearman rank correlation coefficient ρ

$$\rho = 1 - \frac{1}{H}\sum_{h=1}^{H}\frac{6}{21(21^2-1)}\left[\sum_{i=1}^{21}\left(rank_{VERT}(i) - rank_{H_h}(i)\right)^2\right]$$

Experimental results

Table 1. λs with Ranking Weights

	P	R_1	R_{2S}	R_{2D}	User
DATI	0.5317	0.6270	0.5794	0.6270	0.5714
YSL	0.5317	0.7063	0.6905	0.6587	0.6286

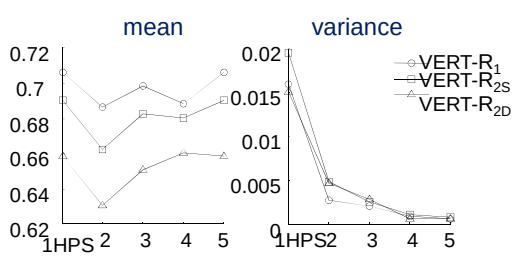
Table 2. ρs with Ranking Weights

	P	R_1	R_{2S}	R_{2D}	User
DATI	0.1071	0.6429	0.4643	0.6429	0.6190
YSL	0.2143	0.7500	0.8571	0.8214	0.6310

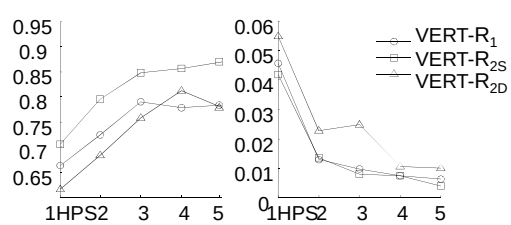
Table 3. λs and ρs with Uniform Weights

	$\lambda(R_1)$	$\lambda(R_{2S})$	$\rho(R_1)$	$\rho(R_{2S})$
DATI	0.6270	0.5794	0.6429	0.4643
YSL	0.6905	0.6905	0.6071	0.8214

Experimental results



Accuracy percentage λ



Spearman rank correlation coefficient ρ

Conclusions

- **VERT-P does not correlate well with human assessment**
  - the values of Spearman coefficients for VERT-P are very small
- **VERT-R measure is effective**
  - the value of APs and Spearman coefficients are both around 0.6
- **Variants of VERT-R have similar performance**
  - Need to extend the experiments in size and scope to further identify the capabilities of the method
- **Future work:**
  - Large scale experiments with Wikio web site

Thank you!

Questions?