Unsupervised Temporal Commonality Discovery

Wen-Sheng Chu, Feng Zhou and Fernando De la Torre

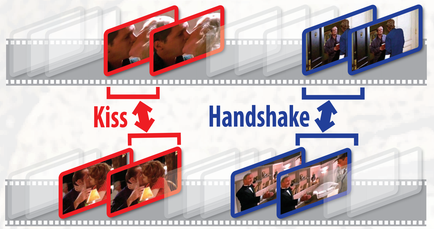
Carnegie Mellon

ECCV 2012

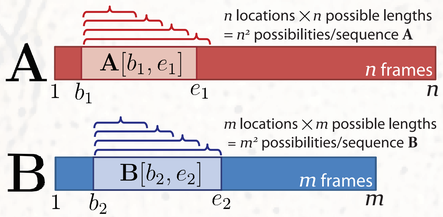
CODE AVAILABLE

NEW Problem Setting

- **Goal**
  - △ Given two videos, discover common events in an **unsupervised** fashion.
  - △ We name this problem **Temporal Commonality Discovery (TCD)**.



- **TCD is difficult**
  1) **No prior knowledge** (what, where, and how many) on commonalities is provided.
  2) Exhaustive search approaches (*e.g.*, sliding window) are **computationally prohibitive**.

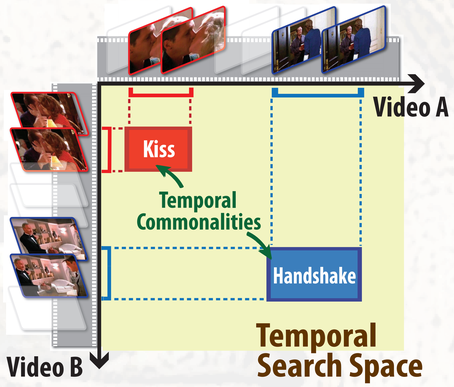


Computational complexity: $O(n^4)$

△ *E.g.*, two videos with 300 frames have **>8,000,000,000** possible matches.

Problem Interpretation

- **Translate TCD into searching 2-D temporal space**
  - △ A rectangle $r=[b_1,e_1,b_2,e_2]$ indicates a solution.
  - △ Search over rectangle sets $R=[B_1,E_1,B_2,E_2]$.



Differences with Previous Work

- Difference between **TCD** and **ESS** [1]/**STBB** [2]
  - △ Different learning framework: **Unsupervised** v.s. **Supervised**
  - △ New bounding functions for **TCD**
- Difference between **TCD** and [3]
  - △ Differenct objective: **Commonality Discovery** v.s. **Temporal Clustering**

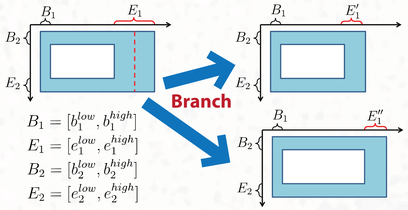
[1] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization", ***PAMI***, pp. 2129–2142, 2009.
[2] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection", ***PAMI***, 33:1728–1743, 2011.
[3] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events", in ***CVPR***, 2010.

The Proposed TCD Method

- Problem Formulation

$$\min_{b_1,e_1,b_2,e_2} d\left(\varphi_{\mathbf{A}[b_1,e_1]}, \varphi_{\mathbf{B}[b_2,e_2]}\right), \quad \text{s.t. } e_i - b_i \geq \ell, \forall i \in \{1,2\}$$

- Optimization by **Branch** and **Bound** (B&B)



- Derive **Bounding functions**

1) Bounding L1 distance:

$$\widehat{l}_b = \max\left(\frac{h_b^-}{|\mathbf{A}^+|}, \frac{k_b^-}{|\mathbf{B}^+|}\right) - \min\left(\frac{h_b^+}{|\mathbf{A}^-|}, \frac{k_b^+}{|\mathbf{B}^-|}\right),$$

$$\widehat{u}_b = \max\left(\frac{h_b^+}{|\mathbf{A}^-|}, \frac{k_b^+}{|\mathbf{B}^-|}\right) - \min\left(\frac{h_b^-}{|\mathbf{A}^+|}, \frac{k_b^-}{|\mathbf{B}^+|}\right)$$

2) Bounding intersection similarity:

$$l_{\cap}(\mathcal{R}) = -\sum_{b=1}^{K} \min\left(\frac{h_b^+}{|\mathbf{A}^-|}, \frac{k_b^+}{|\mathbf{B}^-|}\right)$$

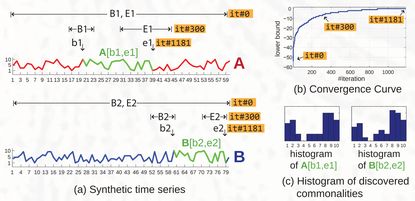
$$u_{\cap}(\mathcal{R}) = -\sum_{b=1}^{K} \min\left(\frac{h_b^-}{|\mathbf{A}^+|}, \frac{k_b^-}{|\mathbf{B}^+|}\right)$$

3) Bounding χ^2-distance:

$$l_{\chi^2}(\mathcal{R}) = \sum_{b=1}^{K} \frac{(\max(0, \widehat{l}_b))^2}{h_b^+/|\mathbf{A}^-| + k_b^+/|\mathbf{B}^-|}$$

$$u_{\chi^2}(\mathcal{R}) = \sum_{b=1}^{K} \frac{\widehat{u}_b^2}{h_b^-/|\mathbf{A}^+| + k_b^-/|\mathbf{B}^+|}$$

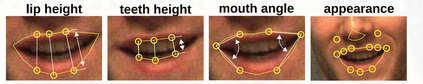
- An example for two synthetic time series:



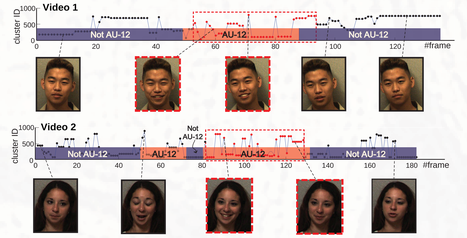
(a) Synthetic time series (b) Convergence Curve (c) Histogram of discovered commonalities

Common Facial Events Discovery

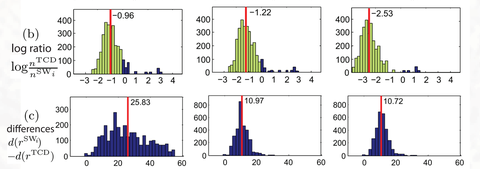
- Evaluation performed on the RU-FACS database [3,4].
- Facial feature representations



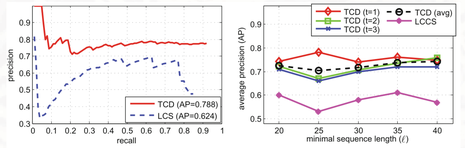
- Example of a dicovered common facial event



- Efficiency evaluation (compared to **Sliding Window**)



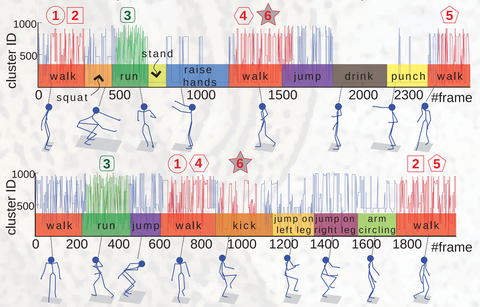
- Performance evalutaion (compared with **LCCS** [5])

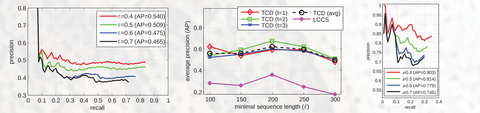


[4] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions", ***Journal of Multimedia***, 1(6), 2006.
[5] Y. Wang and S. Velipasalar, "Frame-level temporal calibration of unsynchronized cameras by using Longest Consecutive Common Subsequence", In ***ICASSP***, 2009.

Multiple Common Motions Discovery

- Evaluation performed on the CMU Mocap [6].





(a) Precision-recall (ℓ1) (b) Average precision (c) Precision-recall (χ^2)

[6] http://mocap.cs.cmu.edu/

Search Video by Video

- Slight modification leads to efficient **video indexing**

$$\min_{b_t,e_t} d\left(\varphi_{\mathbf{T}[b_t,e_t]}, \varphi_{\mathbf{Q}}\right) \text{ s.t. } e_t - b_t \geq \ell.$$

- Example on the Hollywood dataset [7]

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", In ***CVPR***, 2008.