

[招待講演] 映像検索への音声工学からのアプローチ

篠田 浩一^{1,a)}

概要：インターネット上の大量の消費者映像から，そのコンテンツを解析して，情報を抽出する映像意味検索技術が盛んに研究されている．その中で，毎年米国で開催されている TRECVID は，世界の有力研究機関が集まって共通のタスクでの性能を競い合うワークショップで，最先端技術のショーケースとなっている．ここでは，TRECVID タスクのうち，特に意味インデクシングとマルチメディアイベント検出の 2 つに焦点を当て，それらにおいて，音声工学でこれまで培われてきた方法論や開発されてきた特徴量がどのように貢献しているかを解説する．さらに，今後のマルチメディア検索技術の展開を予想する．

インターネット上の映像の中身 (コンテンツ) を解析して，そこから意味の情報を抽出する技術，Content-Based Video Retrieval (CBVR) の研究が盛んに行われている．映像信号の特徴 (低次特徴) と人間にとって意味のある概念 (高次特徴) の間には大きな乖離があり，このセマンティック・ギャップの克服がもっとも大きな課題となっている．

TRECVID [1], [2] は米国の国立標準技術研究所 (NIST) 主催の映像検索のワークショップである．TREC から 2001 年に独立した．共通のタスクを設定してその性能を競うクロードな競争型のワークショップである．IBM, CMU, Columbia 大, アムステルダム大などが参加しており，日本からも NII, NTT, 東工大などが参加している．映像検索のトップクラスの研究者が集まっており，いわば最先端技術のショーケースとなっている．

例えば Semantic INdexing(SIN) は映像のショットから「コンセプト」を抽出するタスクである．ここでのショットとはカメラの切り替わりで区切られる区間を指し，通常数秒～30 秒程度である．コンセプトは，オブジェクト (犬，椅子など)，シーン (夜景，屋外など)，アクション (歌う，踊るなど) である．主な手法は SIFT(Scale Invariant Feature Transform) などの局所特徴を量子化したコードブックを用いる BoW(Bag of Words) である．

また，Multimedia Event Detection (MED) は映像のクリップから「イベント」を検出するタスクである．クリップの長さは 30 秒～3 分程度であり，通常複数のショットから成る．ここで，イベントとは，人間と人間との間や人間から事物への行動を指す．「タイヤを交換している」，「誕

生日を祝っている」などがその例である．従来は，SIN の手法をそのまま適用するアプローチが主流であったが，近年，意味インデクシング，音声認識，OCR などの識別器や入力特徴量のクラスタリングから得られた情報を「中間表現」とし，それらを入力とした検出器を設計するアプローチが試みられている．

音声工学発の技術は主に以下の 3 つの側面で重要な役割を果たしている [3]．

- (1) 多くのコンセプト・イベントを検出するために，音声・音響特徴が画像特徴を補完する役割を果たす．
- (2) 音声で培われた GMM, HMM などの統計・確率的アプローチ，及び，その頑健性を高める様々な手法は，しばしば有効である [4]．また，Deep Learning も今後の展開が期待される [5]．
- (3) リアルタイム動作の必要性から培われた高速化技術が開発効率の向上に貢献している．

謝辞

映像検索の研究についてご支援いただいたキヤノン (株)，及び，研究室の諸氏，特に井上中順氏に感謝する．

参考文献

- [1] <http://trecvid.nist.gov/>
- [2] A. F. Smeaton et al., “Evaluation campaigns and TRECVID”, Proc. MIR’06, 2006.
- [3] K. Shinoda et al., “Reusing speech techniques for video semantic indexing”, IEEE Signal Processing Magazine, vol. 30, no. 2, pp. 118–122.
- [4] N. Inoue et al., “A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM Supervectors”, IEEE Trans. Multimedia, vol. 14, no. 4-2, pp. 1196–1205, 2012.
- [5] C. Snoek et al., “Deep nets for detecting, combining, and localizing concepts in video”, Proc. TRECVID, 2013.

¹ 東京工業大学
Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8552, Japan

^{a)} shinoda@cs.titech.ac.jp