Enhanced Spatio-Temporal Video Copy Detection by Combining Trajectory and Spatial Consistency

Savaş Özkan¹,², Ersin Esen¹, Gözde Bozdağlı Akar²
1:TUBITAK UZAY 2:Middle East Technical University

Overview
Definition
Copyright infringements have become a serious problem;
• Everlasting increase of available video data
• Increase of internet bandwidth availability
• Proliferation of video sharing websites

Content-based video copy detection
• Objective: Obtain sufficient number of signatures from video contents and estimate source of the original video by leveraging these signatures.
• It consists of two steps, namely, offline (extracts signatures and stores in an archive), and online (compares query video signatures within the archive).

Spatio-Temporal Signature
Video Sampling: Initially, trajectory frames are selected in one second intervals. For better estimation, the frames with globally static scenes are discarded.

Trajectory Estimation: To capture foreground motion with high precision, densely sampled trajectory estimation is employed [Wang et al., 2013].
Steps:
• Locate feature points that are densely sampled in multiple scales, s ∈ Z.
• Select good feature points for seamless tracking [Shi et al., 1994].
• Track these points in time and estimate utter trajectories.
Enhancements:
• Filter out the trajectories with large displacements.
• Unlike action recognition, the trajectory with small spatial variation, or implicitly which behaves like spatially static interest point, also have valuable information.

Spatio-Temporal Feature Extraction: Two visual representation methods are computed on each frame separately, HoG [Dolal et al., 2005] and Average Opponent Color [Koen et al., 2010].
Enhancements:
• In order to boost the effectness of the representation, the space-time volume is divided into subcells in spatial and temporal domains as nₓ, nᵧ, n𝑡 ∈ Z

Feature Indexing: To enhance the applicability of the feature pairs search, final spatio-temporal features are represented with small indices using Bag-of-Word [Sivic et al., 2003] indexing procedure and stored with inverted index structure.

Contribution
Motivation
In general, spatially sampled local signatures have shown excellent performance on occlusion clutter and geometric transformation owing to inherent behavior and incorporation of a geometric verification stage [Douze et al., 2010].

However, joint use of spatial and temporal characteristics in signature extraction and verification stages can make the representation more precise.

In order to boost the acceptance of orientation difference is zero, rotation matrix will be identical matrix and 2D transformation can be rewritten as:

\[ m_1^q \times \left[ m_1^q \right]^T \times \left[ a_1^q \right] \]

Thus, \( f_q = m_1^q \times [a_1^q \right] \times [a_1^q \right] + [a_1^q \right] \)

In order to obtain dominant geometric characteristic, scale difference \( s \) and quantized translation difference \( t_q \) ∈ Z are exploited together.

Additionally, term frequency-inverted index frequency (tf-idf) [Sivic et al., 2003] is employed for more accurate signature matching.

Experimental
TRECVID 2009 Content-Based Video Copy Detection
(400 hours of reference videos and 1407 query videos)
• T2: picture-in-picture.
• T3: insertion of pattern.
• T4: strong re-encoding.
• T5: change in gamma.
• T6: decrease quality. (blur, noise, contrast, frame drop)
• T8: post processing. (crop, shift, vertical flip)
• T10: five randomly selected attacks from T2-T8

Time Burden: Induces extra 0.06 second burden for comparison of one second query video with one hundred hours reference videos.

Conclusion
• Use of our novel consistency method in copy detection improves the performances significantly, particularly in T3, T4, T5 and T8 attacks.
• Probably the best accomplished attempt for integration of joint use of visual and temporal contents on video copy detection and retrieval problems.
• The novel verification stage is applicable enough for large video archive.
• The main bottleneck is offline stage is a bit expensive.

References
• J. Shi et al., “Good features to track.”, 1994.