Video Co-summarization: Video Summarization by Visual Co-occurrence

Wen-Sheng Chu[1], Yale Song[2], Alejandro Jaimes[2]
[1]Robotics Institute, Carnegie Mellon University. [2]Yahoo Labs, New York.

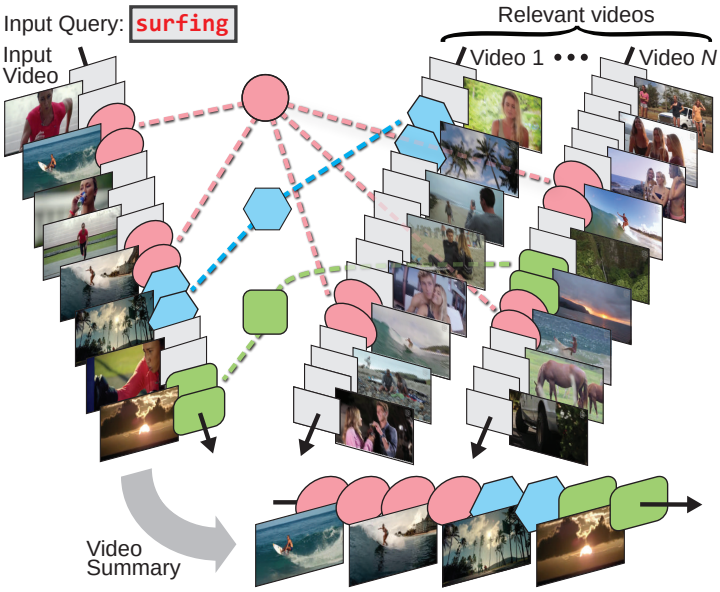

Figure 1: An illustration of **video co-summarization** as identifying visually similar events shared across multiple videos. Different patterns indicate relevant events discovered by our method: surfing (red circles), sunset (green rectangles), and palm tree (blue hexagons).

Figure 2: An illustration of automatic **concept visualization** using our method. Each image indicates a shot on video categories *Bike polo*, *MLB*, *NFL*, *Notre Dame*, *Statue of Liberty*, and *Surfing*.

We present *video co-summarization*, a novel perspective to video summarization that exploits visual co-occurrence across multiple videos. Motivated by the observation that important visual concepts tend to appear repeatedly across videos of the same topic, we propose to summarize a video by finding the most frequently co-occurring shots among videos collected using a topic keyword, as illustrated in Fig. 1.

Formulate co-summarization as finding maximal bicliques: The main technical challenge is dealing with the sparsity of co-occurring patterns, out of hundreds to possibly thousands of irrelevant shots in videos being considered. We develop a Maximal Biclique Finding (MBF) algorithm optimized to find sparsely co-occurring patterns, discarding less co-occurring patterns even if they are dominant in one video.

We model a collection of videos and their associated shots as a weighted bipartite graph. Suppose we are given two videos $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ and $\mathbf{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ with m and n shot-level features, respectively. We model the video pair as a weighted bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = \mathbf{A} \cup \mathbf{B}$ is the vertex set, $\mathcal{E} = \{(\mathbf{a}_i, \mathbf{b}_j) | \mathbf{a}_i \in \mathbf{A}, \mathbf{b}_j \in \mathbf{B}\}$ is the edge set, and $\mathbf{W} = \begin{bmatrix} 0 & \mathbf{C} \\ \mathbf{C}^\top & 0 \end{bmatrix}$ is the weight matrix. Given the graph, we encode the co-occurrence relationship between a pair of videos with a *co-occurrence matrix* $\mathbf{C} \in \mathbb{R}^{|\mathbf{A}| \times |\mathbf{B}|}$. Each entry C_{ij} is computed as the χ^2 distance. Given a set of more than two videos, we apply the same method for each pair of videos to construct the entire graph.

We formulate video co-summarization as finding complete bipartite subgraphs, or *bicliques*. Each biclique represents a compact set of video shots that are visually similar to each other. Specifically, given the co-occurrence matrix $\mathbf{C}$, we look for two binary selection vectors $\mathbf{u}$ and $\mathbf{v}$ that identify the bicliques with maximal visual correlation:

$$\max_{\mathbf{u},\mathbf{v}} \quad \sum_{ij} C_{ij} u_i v_j - \lambda_u \|\mathbf{u}\|_1 - \lambda_v \|\mathbf{v}\|_1 \qquad (1)$$

$$\text{subject to} \quad u_i + v_j \leq 1 + I(C_{ij} \geq \varepsilon), \forall i, j,$$
$$\mathbf{u} \in [0,1]^m, \mathbf{v} \in [0,1]^n,$$

where λ_u and λ_v are tradeoff terms controlling the sparsity in $\mathbf{u}$ and $\mathbf{v}$, and $I(x)$ is an indicator function that returns 1 if the statement x is true, and 0 otherwise. The first constraint ensures that a biclique contains only shots with sufficient visual similarity, *i.e.*, if $C_{ij} < \varepsilon$, either u_i or v_j equals to zero.

Parallelizable with closed-form updates: We derive closed-form updates in (1) using the fact of independent u_i's and v_j's. Let $\widehat{u}_i = \min\{I(C_{ij} \geq \varepsilon) - v_j\}_{j=1}^n$, and $(x)_- = \min(0, x)$ be a non-positive operator, we obtain a closed-form update $u_i = \min\left(I(\mathbf{C}_{i:}\mathbf{v} \geq \lambda_u), 1 + (\widehat{u}_i)_-\right)$. Similarly, $v_j = \min(I(\mathbf{u}^\top \mathbf{C}_{:j} \geq \lambda_v), 1 + (\widehat{v}_j)_-)$, where $\widehat{v}_j = \min\{I(\mathbf{C}_{ij} \geq \varepsilon) - u_i\}_{i=1}^m$. Compared to standard biclique finding algorithms (*e.g.*, [1, 3]), the updates ensure parallelizability with closed-form updates, implying the high scalability of our method. Compared to co-clustering [2] that requires an SVD of $\mathcal{O}(mn^2 + n^3)$, MBF requires only $\mathcal{O}(m+n)$ operations per iteration.

Results: We demonstrate the effectiveness of our approach on motion capture and self-compiled YouTube datasets. Our results suggest that summaries generated by visual co-occurrence tend to match more closely with human generated summaries, when compared to several popular unsupervised techniques. We also show an application of video co-summarization to generating important visual concepts from multiple videos, i.e., we summarize a collection of videos altogether. Fig. 2 illustrates example summaries produced by our approach. A user study performed with 20 participants show the superiority of our approach compared to baselines, suggesting that our approach has successfully generated visualization of important concepts from multiple videos.

[1] Gabriela Alexe, Sorin Alexe, Yves Crama, Stephan Foldes, Peter L Hammer, and Bruno Simeone. Consensus algorithms for the generation of all maximal bicliques. *Discrete Applied Mathematics*, 145(1):11–21, 2004.

[2] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, 2001.

[3] Niranjan Nagarajan and Carl Kingsford. Uncovering genomic reassortments among influenza strains by enumerating maximal bicliques. In *Int'l. Conf. on Bioinfo. and Biomedicine*, 2008.

This is an extended abstract. The full paper is available at the CVF webpage.