Supplementary Material for *Video Co-summarization: Video Summarization by Visual Co-occurrence*

Anonymous CVPR submission

Paper ID 865

Abstract

This supplementary shows more derivation and results that could not be fitted into the main paper due to space limitation. In specific, we provide more details in (1) the derivation of closed-form updates of Algorithm 1, (2) the user study GUI we designed for the experiments of concept visualization, and (3) more results of different methods for visualizing the concept of different video categories.

1. Closed-form updates of Algorithm 1

Suppose $\mathbf{C}$ is the co-occurrence matrix; $\mathbf{u}$ and $\mathbf{v}$ are the two binary vectors we are looking for (see notation[1]). Recall that the sparse maximal biclique problem is written as:

$$\begin{aligned} \max_{\mathbf{u},\mathbf{v}} \quad & \sum_{ij} C_{ij} u_i v_j - \lambda_u \|\mathbf{u}\|_1 - \lambda_v \|\mathbf{v}\|_1 \quad (1) \\ \text{subject to} \quad & u_i + v_j \le 1 + \mathbf{1}(C_{ij} \ge \epsilon), \forall i, j \\ & \mathbf{u} \in [0,1]^m, \mathbf{v} \in [0,1]^n. \end{aligned}$$

Suppose we solve for $\mathbf{u}$ with $\mathbf{v}$ fixed, (1) becomes:

$$\begin{aligned} \max_{\mathbf{u}\in[0,1]^m} \quad & \sum_i (\mathbf{C}_{i:}\mathbf{v} - \lambda_u) u_i \quad (2) \\ \text{subject to} \quad & u_1 \le 1 + \mathbf{1}(C_{1j} \ge \epsilon) - v_j, \forall j, \\ & \vdots \\ & u_m \le 1 + \mathbf{1}(C_{mj} \ge \epsilon) - v_j, \forall j. \end{aligned}$$

As u_i is upper-bounded by n constraints, *i.e.*, $u_i \le 1 + \mathbf{1}(C_{ij} \ge \epsilon) - v_j, \forall j$, for notational simplicity we denote the upper bound as $\widehat{u}_i = \min\{\mathbf{1}(C_{ij} \ge \epsilon) - v_j\}_{j=1}^n$. Because u_i are independent of each other, we can derive the update for each entry of $\mathbf{u}$ as:

$$u_i = \begin{cases} 0, & \text{if } \mathbf{C}_{i:}\mathbf{v} < \lambda_u, \\ \min(1, 1+\widehat{u}) = 1 + \min(0, \widehat{u}), & \text{otherwise.} \end{cases} \quad (3)$$

Introducing a non-positive operator $(x)_- = \min(0, x)$, we can rewrite the second case in (3) as $u_i = 1 + (\widehat{u}_i)$, and rewrite the update (3) in a more compact way:

$$u_i = \min\left(\mathbf{1}(\mathbf{C}_{i:}\mathbf{v} \ge \lambda_u), 1 + (\widehat{u}_i)_-\right). \quad (4)$$

Similarly, we can write the upper bound for v_j as $\widehat{v}_j = \min\{\mathbf{1}(C_{ij} \ge \epsilon) - u_i\}_{i=1}^m$, and its update in compact way:

$$v_j = \min\left(\mathbf{1}(\mathbf{u}^\top \mathbf{C}_{:j} \ge \lambda_v), 1 + (\widehat{v}_j)_-\right). \quad (5)$$

Since each element in $\mathbf{u}$ and $\mathbf{v}$ can be computed independently, these closed-form updates are parallelizable, implying our proposed algorithm can be extended larger scale dataset.

2. User study GUI for concept visualization

For the concept visualization experiment in Sec. 4.3 of the main manuscript, we designed a user study by building a AMT-like webpage. Figs. 1 and 2 illustrate the snapshots of the evaluation webpage. The goal is to identify how human perceive video summaries give the videos sharing one topic, *e.g.*, *Surfing*. Before the study begins, we provide the introductory paragraph as follows:

> "*Imagine a video search website that shows you a quick summary of each video as a preview. The quality of video summaries (e.g., relevance to the query/title/etc.) is crucial for better user experience.*
>
> *We want to understand how people perceive the quality of video summaries according to a specific query. This questionnaire will gather data to help answer this question.*"

[1] Bold capital letters denote a matrix $\mathbf{X}$; bold lower-case letters denote a column vector $\mathbf{x}$. $\mathbf{X}_{i:}$ and $\mathbf{X}_{:j}$ represent the i-th row and the j-th column of the matrix $\mathbf{X}$, respectively. $\mathbf{e}_n$ denotes an n-dimensional vector of ones. All non-bold letters represent scalars. X_{ij} denotes the scalar at the (i,j) element of matrix $\mathbf{X}$. x_i denotes the scalar in the i-th element of $\mathbf{x}$.

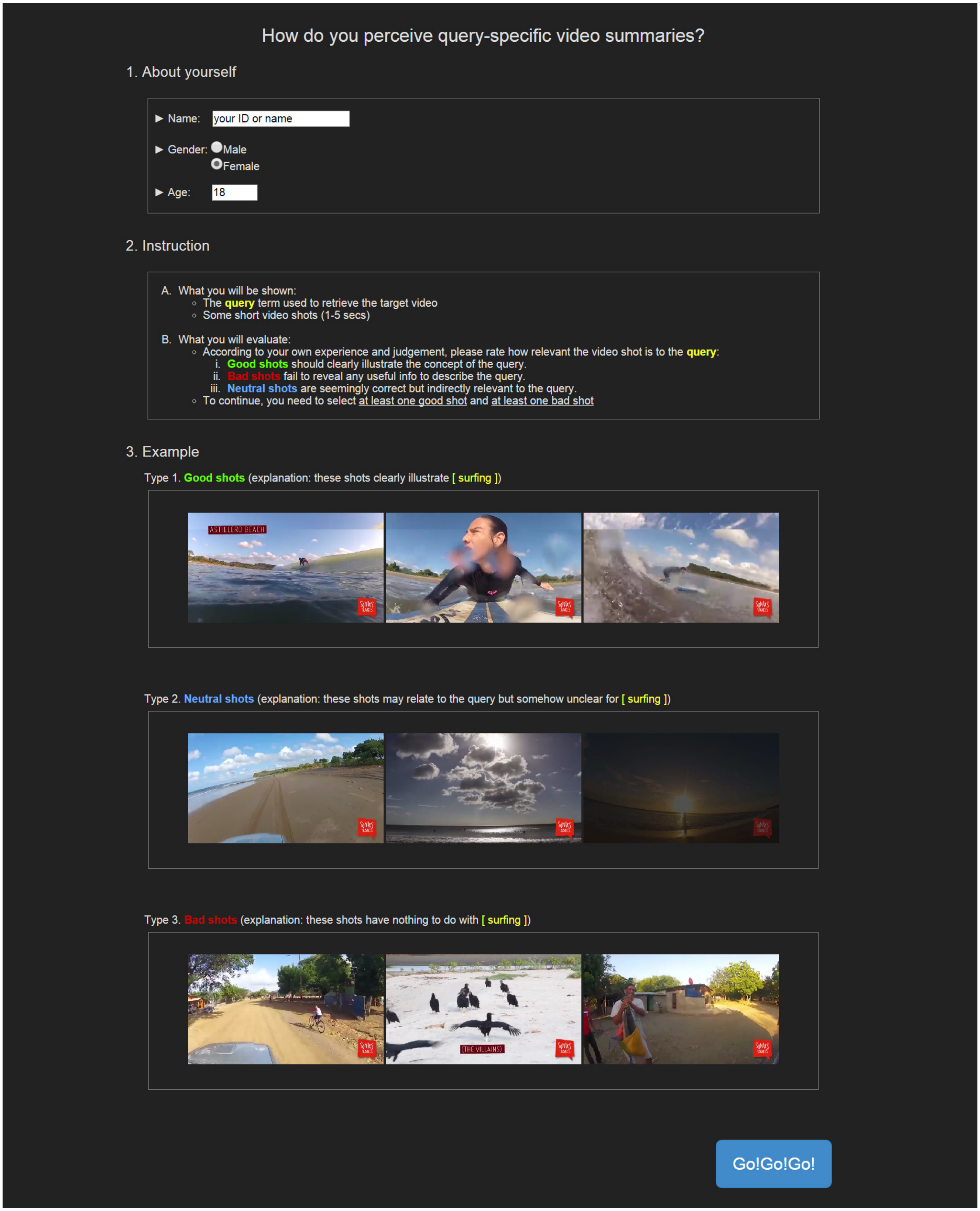

Figure 1. Entrance page of our AMT-like webpage for user study on concept visualization in Sec. 4.3 of the main manuscript. Part 1 collects the user demography. Part 2 gives instructions about what the users will evaluate. Part 3 provides examples about Good, Neutral and Bad summaries using the query *Surfing*.

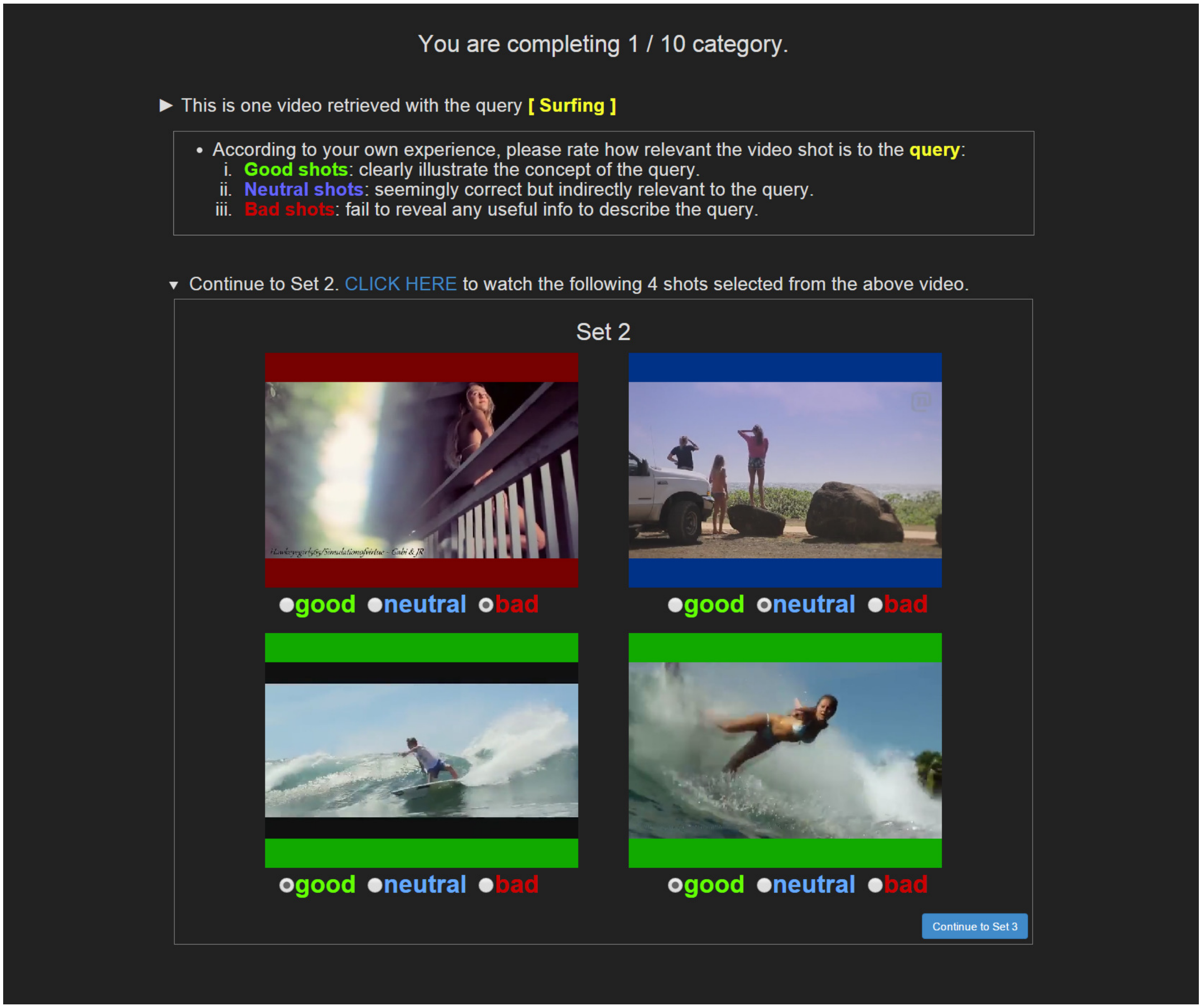

Figure 2. Evaluation page of our AWT-like webpage for user study on concept visualization in Sec. 4.3 of the main manuscript. The progress, *i.e.*, 1 out of 10 categories, in shown on top of the webpage. The query string *Surfing* is highlighted with bright yellow color. A quadruplet of summaries (each generated from different methods) is shown at a time. The order of the summaries is completely random. Totally 5 sets of quadruplets are shown to a user. A user is required to select at least one good summary and one bad summary to continue to the next set of videos. To assist the users, we used green, blue and red colors to indicate Good, Neutral and Bad summaries.

To perform a fair comparison, we provided a quadruplet of 4 summaries at a time (as described in Sec. 4.3 in the manuscript). Each summary was generated from one method, *i.e.*, k-means (baseline approach), LiveLight [2] (state-of-the-art unsupervised video summarization), co-clustering [1] (originally for word-document classification and our first attempt), and MBF (the proposed maximal biclique finding algorithm). The order of the summaries were randomized in a way that the viewers cannot tell by which method each summary was generated, and have to select at least one good summary and one bad summary to continue to the next quadruplet. This study was carried out across 20 subjects (14 males and 6 females ranging from 23 to 33 years old). The entire evaluation took about 30 minutes for a user.

3. More results on concept visualization

This section shows the qualitative results of concept visualization that could not be fitted into the main manuscript due to space limitation. Figs. 3∼12 show the results of different methods on visualizing the concept of each video category. Each row shows the results of k-means (baseline approach), LiveLight [2] (state-of-the-art unsupervised video summarization), co-clustering [1] (originally

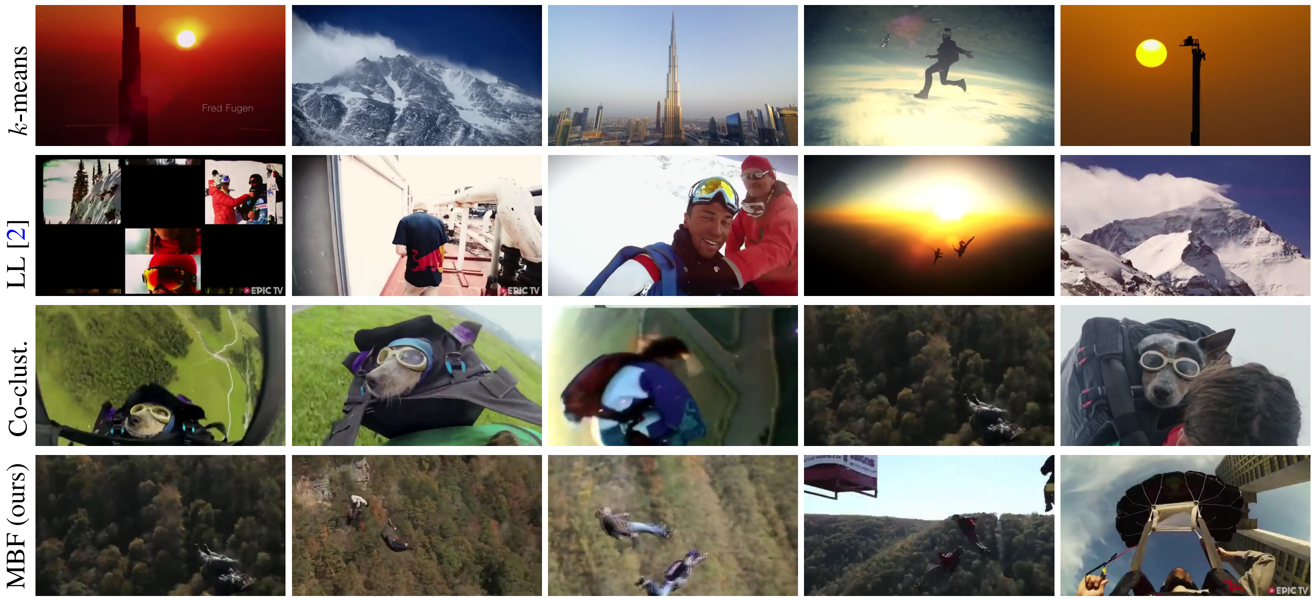

Figure 3. Illustration of different methods on visualizing the concept of *Base jump*.

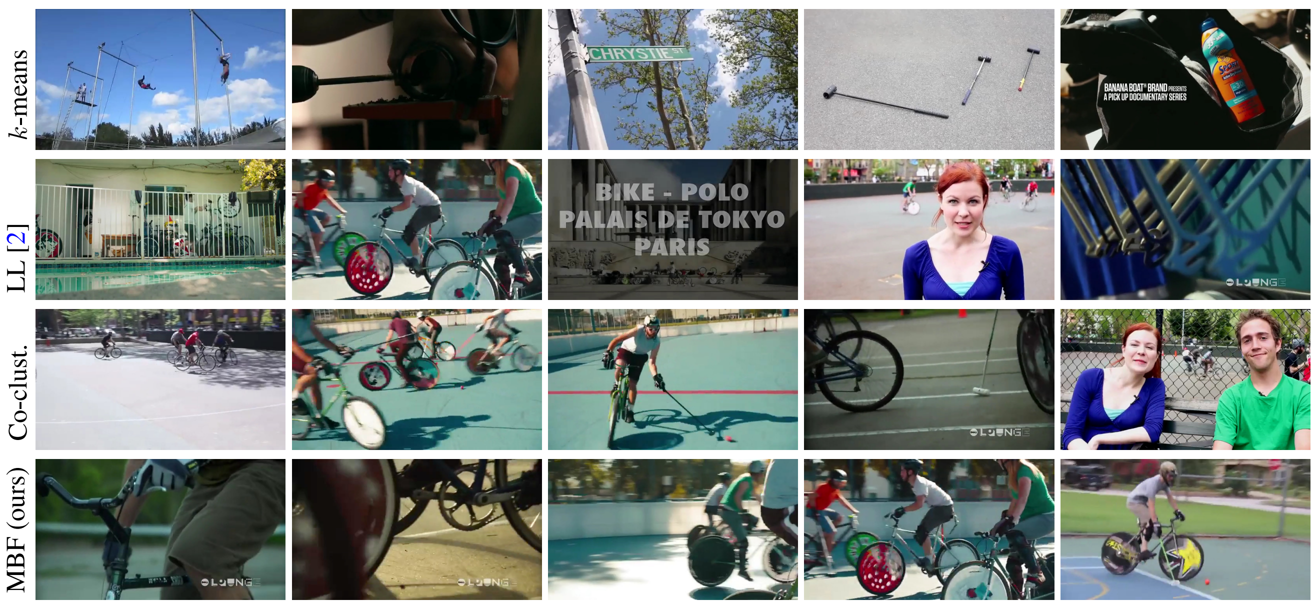

Figure 4. Illustration of different methods on visualizing the concept of *Bike polo*.

for word-document classification and our first attempt), and MBF (the proposed maximal biclique finding algorithm), respectively. In each row, the summaries are ranked from left to right according to the quality scores computed in Sec. 3.3 in the manuscript. Each summary is represented as a snapshot that was taken as the central frame of a summary.

References

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

[2] B. Zhao and E. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014.

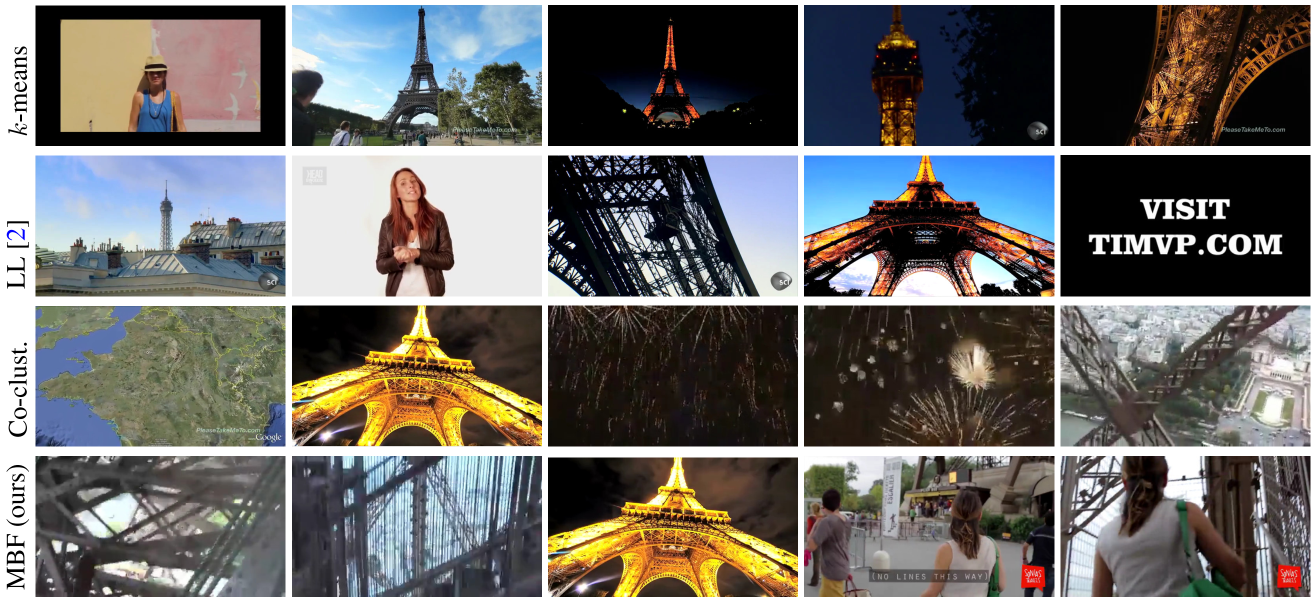

Figure 5. Illustration of different methods on visualizing the concept of *Eiffel Tower*.

Figure 6. Illustration of different methods on visualizing the concept of *Excavators river Crossing*.

Figure 7. Illustration of different methods on visualizing the concept of *Kids playing in leaves*.

Figure 8. Illustration of different methods on visualizing the concept of *MLB*.

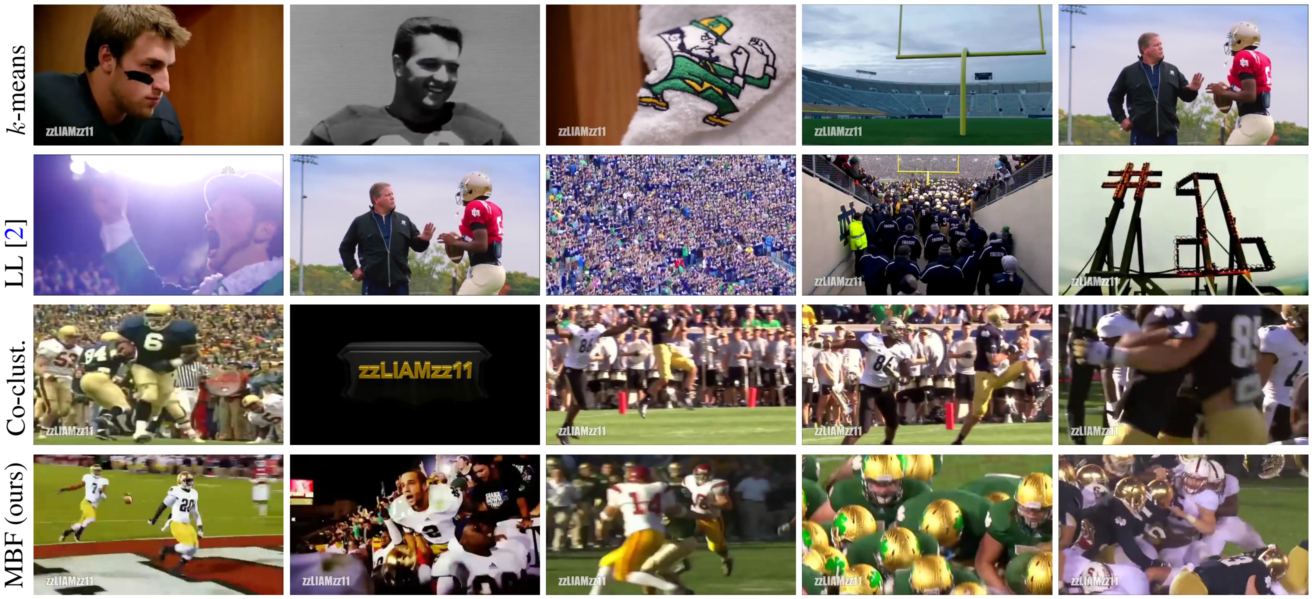

Figure 9. Illustration of different methods on visualizing the concept of *NFL*.

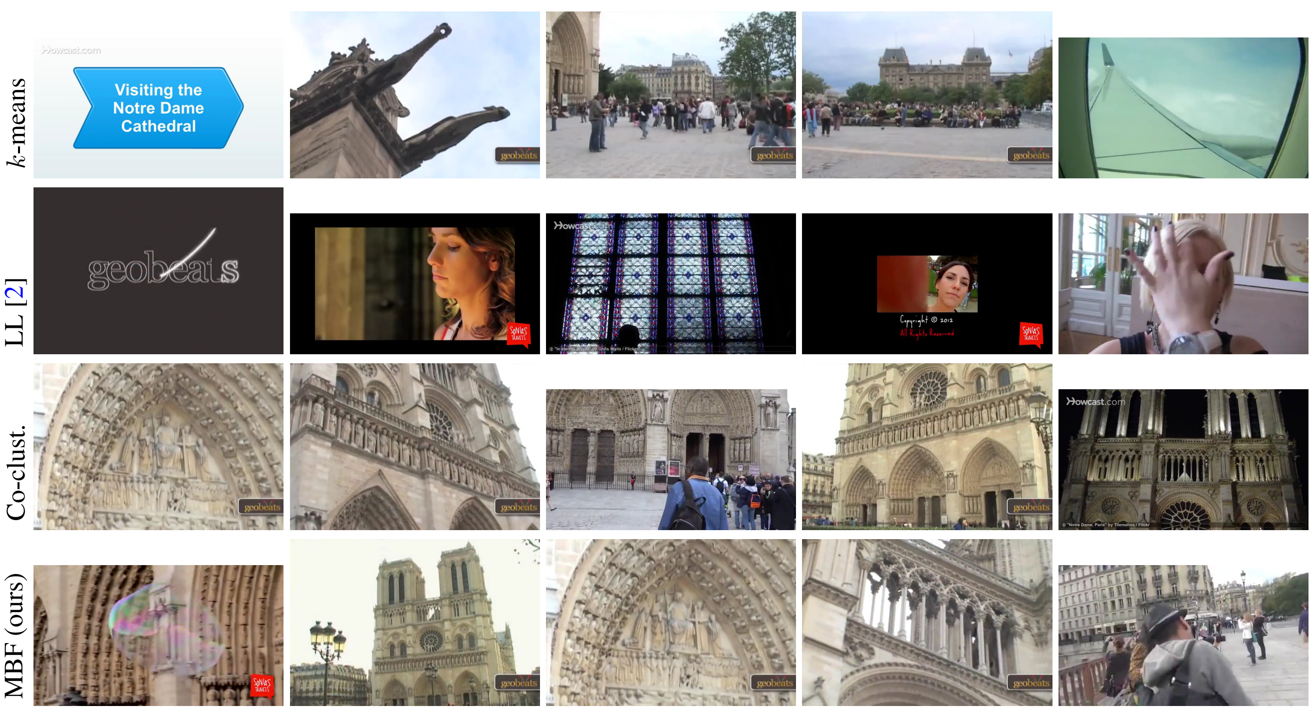

Figure 10. Illustration of different methods on visualizing the concept of *Notre Dame Cathedral*.

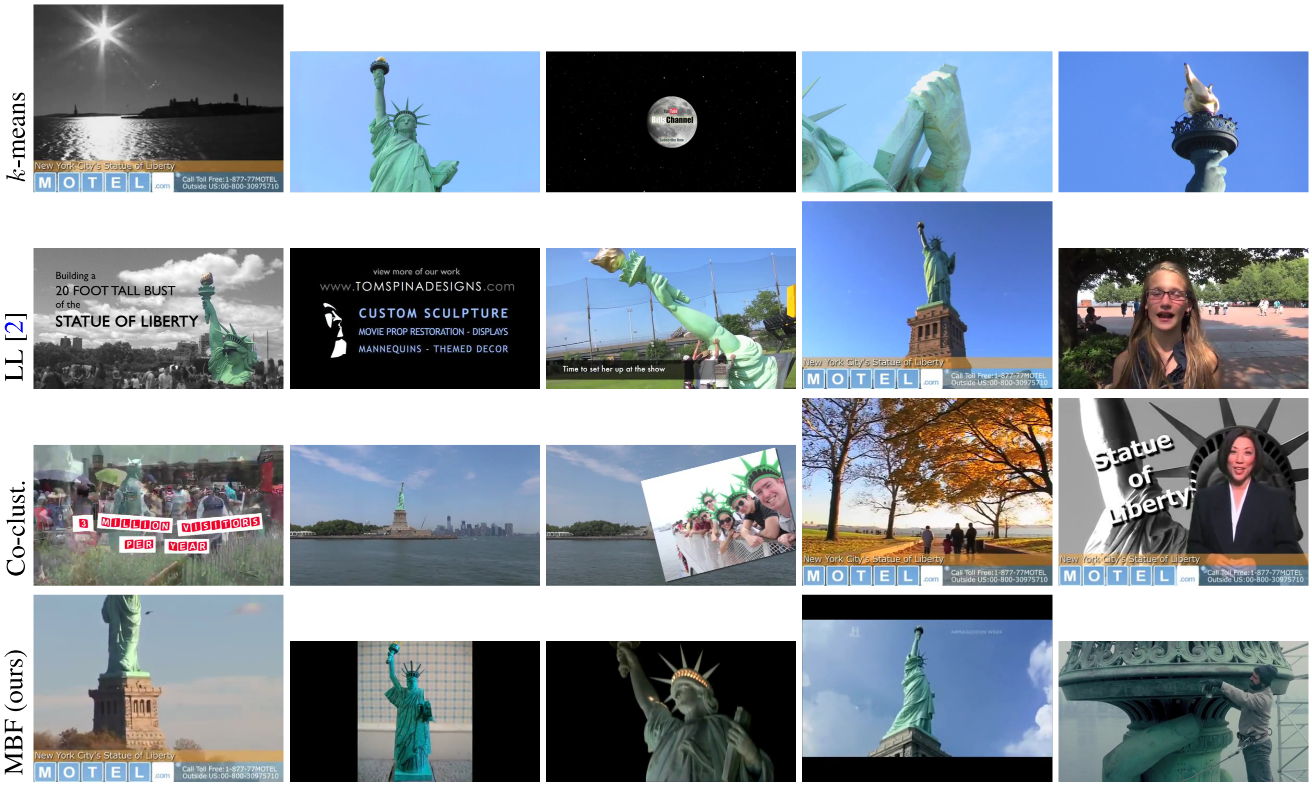

Figure 11. Illustration of different methods on visualizing the concept of *Statue of Liberty*.

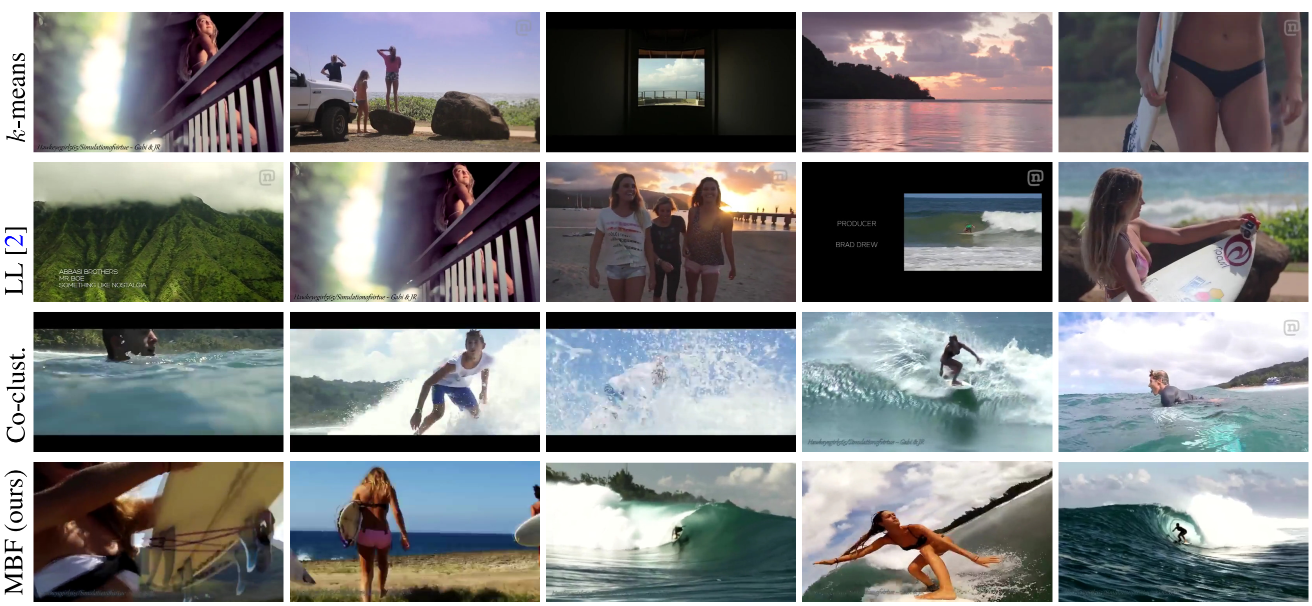

Figure 12. Illustration of different methods on visualizing the concept of *Surfing*.