The task of object mining has attracted a lot of attention of vision researchers. An efficient object mining scheme has many applications, including a mechanism for accessing data in an organized manner, selective video playback and summarization. Traditionally this task has been attempted in two ways. Several approaches cast this problem as one of extracting frequently occurring patterns. The over-all problem is then cast in the manner analogous to searching in a text corpus [7], [8], [5]. Such approaches fail to utilize temporal redundancy in video data. Another set of approaches attempt to organize a very large number of images in an ordered set through object mining [1], [4]. However, they either suffer from low recall rate or provide a very redundant representation.

In the proposed work, we perform object grouping to capture temporal redundancy in video data and provide a compact graph-based representation. Further, we utilize scripts and subtitles to construct a separate graph-based representation for each scene. Advantages of the proposed scheme include faster object retrieval, efficiency in further processing of video data and summarization as a spin-off.

Scripts and subtitles provide complimentary information for an underlying video. For combining such complimentary information robustly, we utilize Lavershtein distance with dynamic time warping. This process is shown in Figure 1.

Scene information from this combined information is then utilized to divide data into the number of distinct scenes. Further, video shot detection is performed by using color histogram based technique and a representative frame is selected for each video shot. In order to capture redundancy across frames of a video shot, we perform object grouping by detecting and tracking maximally stable [2] and Harris affine regions [3] across frames of a video shot. We perform object grouping by detecting and tracking maximally stable [2] and Harris affine regions [3] across frames of a video shot. We compute distance between two such trajectories, $d(T_i, T_j)$, in the following manner:

$$d(T_i, T_j) = \begin{cases} \frac{1}{|T_i| |T_j|} \sum_{T_k \in T_i, T_l \in T_j} d^k(x_i, x_l) \times d^k(h_i, h_l); & \text{if } |T_i \cap T_j| > 0 \\ 0; & \text{Otherwise} \end{cases}$$

where $d^k(T_i, T_j)$, $d_k$, $d^k$, $x$ and $h$ denote distance between trajectories, $k^{th}$ frame of a video shot, spatial distance between trajectories, distance in appearance of trajectories, spatial coordinates of a trajectory and local color histogram of a trajectory, respectively. Spatial distance is taken to be an $L_2$ norm of a difference between spatial coordinates and appearance distance is an $L_1$ norm of a distance between local color histograms. Similarity matrix, $S$, for a shot is computed in the following manner:

$$S(i, j) = e^{\frac{-d^k(T_i, T_j)}{\sigma^2}}; \quad \sigma^2 = \frac{1}{\sum_{i, j} |T_i \cap T_j| > 0} \sum_{i, j} \sum_{T_k \in T_i, T_l \in T_j} \frac{d^k(x_i, x_l)}{|T_i \cap T_j|}$$

where $|T_i|$ is an indicator function. We decide upon number of objects present across a video shot through the following:

$$\text{Number of clusters} (c) = \sum_k \left[ \lambda_{\max}^k / \lambda_k > \theta \right]$$

where $\lambda_{\max}$ and $\lambda_k$ are maximum and $k^{th}$ eigenvalues, respectively. $\theta$ was empirically selected to be 15. Subsequently, we utilize normalized cuts [6] to generate object definitions. Each object definition was taken from a representative frame and consists of visual vocabulary labels and their spatial coordinates, $L$ and $X$, respectively. Other instances of an object occurring in frames other than a representative frame of a video shot serve as attributes to this vertex, formed by an object definition from a representative frame. These attributes are object instances, $O$, and point matches, $M$, respectively. Intermittently occurring object instances are also stored as attributes to the representative object definition by performing threading across video shots. Example graph-based representation for a scene containing three video shots is shown in Figure 2 along with the typical vertex representation. We represent each object definition through the tf-idf descriptor, $W$. Finally, object definitions are connected through the edges of a graph, where each edge represents a similarity between two object definitions, computed through similarity between tf-idf descriptors and spatial re-ranking. In manuscript, we show that the representation obtained from the proposed scheme is much more compact than other state-of-the-art methods, which enables faster object retrieval. The proposed representation also yields video summarization as a spin-off. Finally, we conclude that the proposed mechanism provides comparable precision and higher recall compared to the approach of [7] with faster retrieval of object instances and summarization of a video.