2009 Seventh International Workshop on Content-Based Multimedia Indexing

Hierarchical Summarisation of Video using Ant-Tree Strategy

Tomas Piatrik
Multimedia and Vision Research Group
Queen Mary, University of London
London E1 4NS, UK
tomas.piatrik@elec.qmul.ac.uk

Ebroul Izquierdo
Multimedia and Vision Research Group
Queen Mary, University of London
London E1 4NS, UK
ebroul.izquierdo@elec.qmul.ac.uk

Abstract

Video summarisation approaches have various fields of application, specifically related to organising, browsing and accessing large video databases. In this paper, the appropriateness of biologically inspired models to tackle these problems is discussed and suitable strategy for unsupervised video summarisation is derived. In our proposal, we model the ability of ants to build live structures with their bodies in order to discover, in a distributed and unsupervised way, a tree-structured organization and summarisation of the video data. An experimental evaluation validating the feasibility and the robustness of this novel approach is presented.

1. Introduction

With new video material being created every day, the users need to spend more and more time viewing the content even though they might not be interested into all its aspects. Therefore efficient browsing and access to relevant video items is a crucial application in modern multimedia systems. Video on demand, news broadcasting and syndication, personal video archiving, surveillance and event detection are just few examples of many important multimedia applications in which automatic content structuring is critically needed.

Looking at the video from bottom to top, essential part of a video are frames, which are further organised into shots [3][19]. The next level in the video structure is a scene, defined as a group of shots with similar semantic content. Shots can still be seen as low level of video organization, while scenes represent more semantic level of video organization. After the scenes are found one need to choose which part of the scene actually contains the information important to the user and present it. In this context, two basic image processing tasks, namely, clustering and summarisation play a decisive role.

Research efforts put in the problem of video summarisation and scene detection resulted in large amount of available literature on this topic. Most of the approaches focuses on either detecting the scenes within the video or on defining some importance measure which is the used for selecting parts of the video which will be presented to the user. Main problem in both approaches is still to connect the low level representation of the video with high level semantic representation. Automatic detection of informative parts of a video can be very hard task having in mind subjective nature of importance definition. In [10] authors presented the tool that utilises MPEG-7 visual descriptors and generates a video index for summary creation. The resulting index generates a preview of the movie and allows non-linear access to the content. This approach is based on hierarchical clustering and merging of shot segments that have similar features and neighboring them in the time domain. In [14] Rasheed and Shah construct a shot similarity graph, and use graph partitioning normalised cut for clustering shots into scenes. Video summarisation based on the optimization of viewing time, frame skipping and bit rate constraint is described in [11]. For a given temporal rate constraint the optimal video summary problem is defined as finding a predefined number of frames that minimise the temporal distortion. Otsuka et. al in [12] presented their video browsing system that uses audio to detect sport highlights by identifying segments with mixture of the commentators excite speech and cheering. Video motion analysis can be used for creating video summaries as in [18]. In this approach Wang et al. showed that by analysing global/camera motion and object motion is possible to extract useful information about the video structure. Agglomerative hierarchical clustering with time constraint has been used for shot clustering in [17]. This approach can produce the tree-structured representation that is useful for video summarisation.

More recently, novel bio-inspired approaches have been successfully applied as alternative methods for clustering [2][4][9][15]. In this work we investigate the application of new biological model based on the behavior of real ants for clustering of visual information and summarisation of video content. The rest of the paper is organised as follows. In the next section, an introduction to an ant colony inspired techniques with focus on Ant-Tree Strategy (ATS) is given. The proposed ATS approach for video summarisation and scene detection is presented in Section 3. Selected results of a comprehensive evaluation of the introduced method using relevant video datasets are presented in Section 4. The paper

978-0-7695-3662-0/09 $25.00 © 2009 IEEE
DOI 10.1109/CBMI.2009.50

closes with relevant conclusion and an outlook to future work in Section 5.

2. Ant-Tree Strategy for Clustering

Recent research aimed at improving machine learning capabilities have been strongly influenced and inspired by natural and biological systems [4][7][8]. No one doubts on the efficiency of such systems, optimized over millions of years of natural selection, to maintain and improve life and deliver optimal solution to the extremely complex problems in biological colonies and societies. Relevant studies conclude that the self-organization of neurons into brain-like structures, and the self-organization of ants into warms are similar in many respects [13]. An ant colonies can be seen as "intelligent entities" for its great level of self-organization and the complexity of the tasks they can accomplish. Their colony structure and cooperative behavior inspired many researchers in the field of Computer Science to develop new solutions for optimization problems. For that reason, many novel optimization and clustering algorithms based on the behavior of real ants can be found in the literature, e.g., ACO [4], Ant-Class [15], Ant-Clust [9], and Ant-Tree [2].

Figure 1. Ants filling the gap on the leaf.

The work presented in this paper builds on generic Ant-Tree Strategy, inspired by self-assembling behavior of ants and their ability to build mechanical structures. These types of self-assembly behavior have been observed with Linepithema humiles Argentina ants and African ants of gender Oecophylla longinoda [16]. These colonies of ants exhibit a clear size distribution of workers and they also vary in color from reddish to yellowish brown dependent on the species. Biologists observed the ability to build "chains of ants" in order to fill a gap between two points, or build a nest by closing the edges of a leaf (Figure 1). Interesting observation is that these chains of ants are built according to the similarity between ants. From these self-assembly behaviours, properties that constitute the framework of ATS algorithm can be extracted:

- ants build this type of live structures starting from a fixed support (stem, leaf,...),
- ants can move on this structure whilst it is being currently built. Direction of the next movement is defined by an incoming link from other ants and number of an outgoing links towards other ants,
- ants carry feature vectors containing characteristic information about which of the partitions they should belong to,
- ants maintain and update threshold values for similarity and dissimilarity.

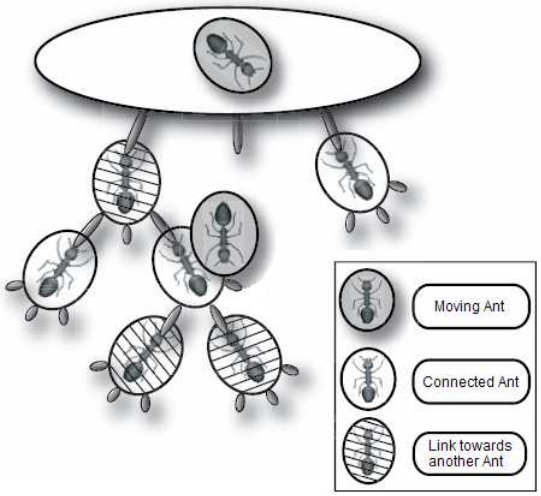

Figure 2. Building of a tree with artificial ants.

The main principles of the algorithm are depicted in Fig. 2, where ants represent a nodes of the tree to be assembled. Each ant can be described by some representation provided that there exists a similarity measure $sim(i, j)$ between two ants $i \in [1, N]$ and $j \in [1, N]$, where N indicates the number of ants. The movement and fixing of ants in a position depends on the similarity value and the local neighbourhood of the moving ants. During the tree construction, ants may fail to connect themselves to the tree. In case of such failure, the updating of threshold allows the ants to be more tolerant and increases their probability of being connected to the tree during the next iteration. Following are the similarity and dissimilarity updating rules:

$$T_{sim}(i) = T_{sim}(i) \cdot 0.9, \tag{1}$$

$$T_{dissim}(i) = T_{dissim}(i) + 0.01, \tag{2}$$

where i represents the ant whose threshold values are being updated. The values for $T_{sim}(i)$ and $T_{dissim}(i)$ are initialised as 1 and 0 respectively.

Each ant starting from root of the tree examines its similarity with the ants that are already connected to the root. If no ants are connected to the root, then the ant connects itself to the root. If the root has some ants already connected

to it, then the current ant finds the most similar ant connected to the root. Following rules are used to define similar and dissimilar ants:

1) If $sim(i,j) > T_{sim}(i)$, then ant i is similar to ant j.
2) If $sim(i,j) < T_{dissim}(i)$, then ant i is dissimilar to ant j.
3) If $sim(i,j) \geq T_{dissim}(i)$ and $sim(i,j) \leq T_{sim}(i)$, then ant i is neither similar nor dissimilar to ant j and needs further processing.

In the above rules, i represents the current ant and j represents the most similar ant connected to the root. If ant i ends up being dissimilar to ant j, then it attaches itself to the root. In the second case, ant i moves itself in the direction of ant j if it ends up being similar and changes it position from root to the ant j. During the next iteration, ant i examines its similarity with all ants connected to ant j and repeats the same process till it finds a suitable place. If it ends up being dissimilar to all ants connected to the current location, then it moves itself in the direction of the most similar ant. One should notice that the next movement of ant i is only limited to incoming and outgoing links from current position. For the third case, if ant i is neither similar nor dissimilar to the ant j, then it updates its similarity and dissimilarity threshold values using (1) and (2). While ant i updates its threshold values, other ants can proceed to build the tree. The algorithm ends when all ants are connected.

3. Video Summarisation using Ants

In this section, a novel video summarisation approach based on ATS is presented. We model the ability of ants to build live structures with their bodies [24] in order to discover, in a distributed and unsupervised way, a tree-structured organization and summarisation of the video data. To be independent from any type and genre of video, we avoid any supervised learning. Ant-Tree clustering algorithm produces very accurate results without using any previous information of the possible distribution of the data set.

In the first step of our algorithm, the original video is divided into the frames and low-level features are extracted from each video frame. For speeding up we can only process every k^{th} frame from the video. Assuming that we get N frames, the inter-similarity matrix $W[N \times N]$ is made of pair wise similarities $sim(i,j)$ between frames $i \in [1,N]$ and $j \in [1,N]$. In the second step, the video frames are clustered using Ant-Tree strategies where each ant represents one video frame. Each frame is attached to the tree structure according to the similarity $sim(i,j)$ and two thresholds for similarity and dissimilarity, which are locally updated by each processed video frame according to (1) and (2). To obtain a video summary, the tree is built so that nodes correspond to video frames and the edges are what needs to be discovered in order to summarise the content (see Figure 3).

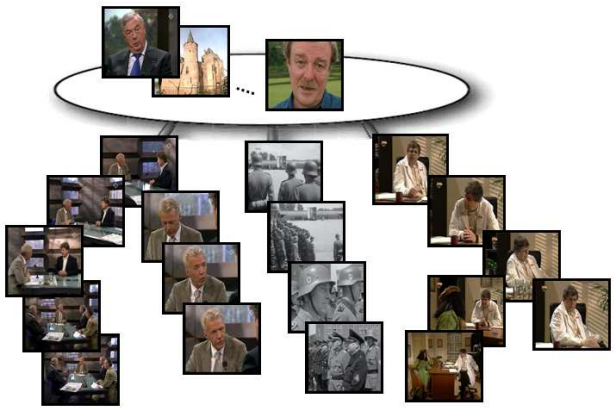

Figure 3. Tree structuring of video frames by Ant-Tree Strategy.

In order to improve the quality of clustering results, the following procedure is implemented. Frames are able of changing a previous assigned location in the tree structure and consequently they can be reallocated in an other, more adequate group. This procedure is inspired by real ants behaviour in formation of drops, where each ant can be detached from the tree or whole drop can sometimes fall down [16]. This iterative process considers each frame in the tree structure be evaluated in order to deciding if it should be disconnected from the tree or not. This decision is based on the Silhouette function [1] which estimates the membership degree of an arbitrary frame with respect to the group under consideration. The Silhouette function is defined as follows:

$$S(i) = \frac{\Lambda_{min}(i) - \Gamma(i)}{max(\Gamma(i), \Lambda_{min}(i))}, \quad (3)$$

where $\Gamma(i)$ is the distance between i^{th} frame and the mean value of the group that i^{th} frame belongs, $\Lambda_{min}(i)$ is the minimum distance between frame i and the mean values of the remaining groups, and $-1 \leq S(i) \leq 1$. A threshold $T = T_0$ must be defined to detect the assignation of i^{th} frame to a wrong group, i.e., when $S(i) < T$. The reallocation process is carried out by a simple movement of the frame toward the next most similar group.

Results of the Ant-Tree algorithm are clusters which are used in the decision process for creating video summaries. It might happen that some of the video frames going consecutively in time are clustered to different clusters. To make algorithm less sensitive to such a small errors, we define the threshold for the maximum number of negative clustered frames going consecutively. If the number of negative clustered frames is above this threshold, the cut is detected in the video scene. In order to present most important information of video content to the user in short time, the following procedure is applied. We defined certain scenes of no interest to the user according to the type of video. In the news domain it is the scene showing an

Table 1. Results for the summarisation of surveillance videos.

$Method$	$detected\ events$	$missed\ events$	$total\ lenght\ of\ summary[min]$
ATS	23	8	$1,8$
SC	25	6	$2,2$
HAC	18	13	2

Table 2. Results for the summarisation of news videos.

$Method$	$detected\ events$	$missed\ events$	$total\ lenght\ of\ summary[min]$
ATS	26	3	3
SC	23	6	3
HAC	19	10	2

"anchor person" and in the surveillance domain statical "indoor" and "outdoor" scene. After the video segmentation is done and the above mentioned scenes are recognised we excluded clusters contained these scenes from the summary using constraints. The final summary is built by putting together one video segment from each cluster that contains the representative frame (first frame of the subtree).

4. Experimental Evaluation

In this section we evaluate the proposed algorithm using video datasets from different domains. We use the combination of MPEG-7 color layout and edge histogram features for representing each selected frame. First, we compare proposed ATS algorithm with Spectral clustering (SC) and Hierarchical Agglomerative clustering (HAC) methods in the scene detection task using news and surveillance videos. Both methods have been successfully used for summarisation of video content and they were implemented in a same fashion as in [5] and [17]. The obtained results are compared with a manually created ground truth of semantic events appeared in videos. Additionally, the proposed method has been implemented as a part of the collaborative video summarisation system and tested on TRECVID 2008 BBC rushes videos.

4.1. Comparative evaluation on videos from surveillance and news domain

For experimental evaluation in the surveillance domain we used two 40 minutes long videos (one depicting an inside room with people entering and leaving and the other depicting an outdoor parking), and in the news domain we used two 30 minutes long videos. Low-level features are computed on the frame level using one frame per second sample ratio. We depicted part of the video (see Figure 4) to show the performance of ATS algorithm in the surveillance domain. Summarisation results of tested algorithms are shown in Tables 1 and 2. The spectral method gives slightly better performances in case of the surveillance domain (80% of events are detected) since it can detect regions with high motion activity, while the ATS method detects 74% of events. The video segments included in the summary of ATS method are shorter comparing to the other methods. In the news domain where anchor person precedes most of the stories the scene recognition properties of the ATS method lead to higher performances 89% of correctly detected event comparing to 79% for the SC and 65% for the HAC approach.

4.2. Trecvid 2008 BBC Rushes evaluation

We have applied proposed algorithm as a part of the collaborative system for automatic video summarisation. This video summarisation system has been developed by the partners of the K-Space European Network of Excellence for the TRECVID 2008 BBC rushes summarisation evaluation. The system is organised in several steps. First, the common segmentation of the video is built, secondly the common segments are evaluated for redundancy and relevance, and finally, a video summary is constructed by concatenating the video clips of the selected segments with a video acceleration. Our approach analysed the common segments to detect redundancies and assess relevance. We use two strategies for determining segments to be included into the summary. One is the explicit selection of segments that are found to be relevant. For each of these segments, a relevance value is determined. The other is to determine redundant segments that shall not be included. The redundancy of a segment can be absolute (i.e. the content is not needed, e.g. a shot containing a color bar) or relative with respect to a set of segments, i.e. these segments contain the same content and only one out of such a set needs to be considered. The Trecvid 2008 rushes dataset consist of 39 videos with total length around 20 hours. Results of the ant-tree algorithm are clusters which are used in the decision process for classification of relevant/redundant segments. Common video segments which contain representative frames attached to the root of the tree are classified as relevant. The importance of relevant segment is defined by number of frames from redundant segments in the corresponding cluster. For selection of redundant segments, we detect monochrome parts of

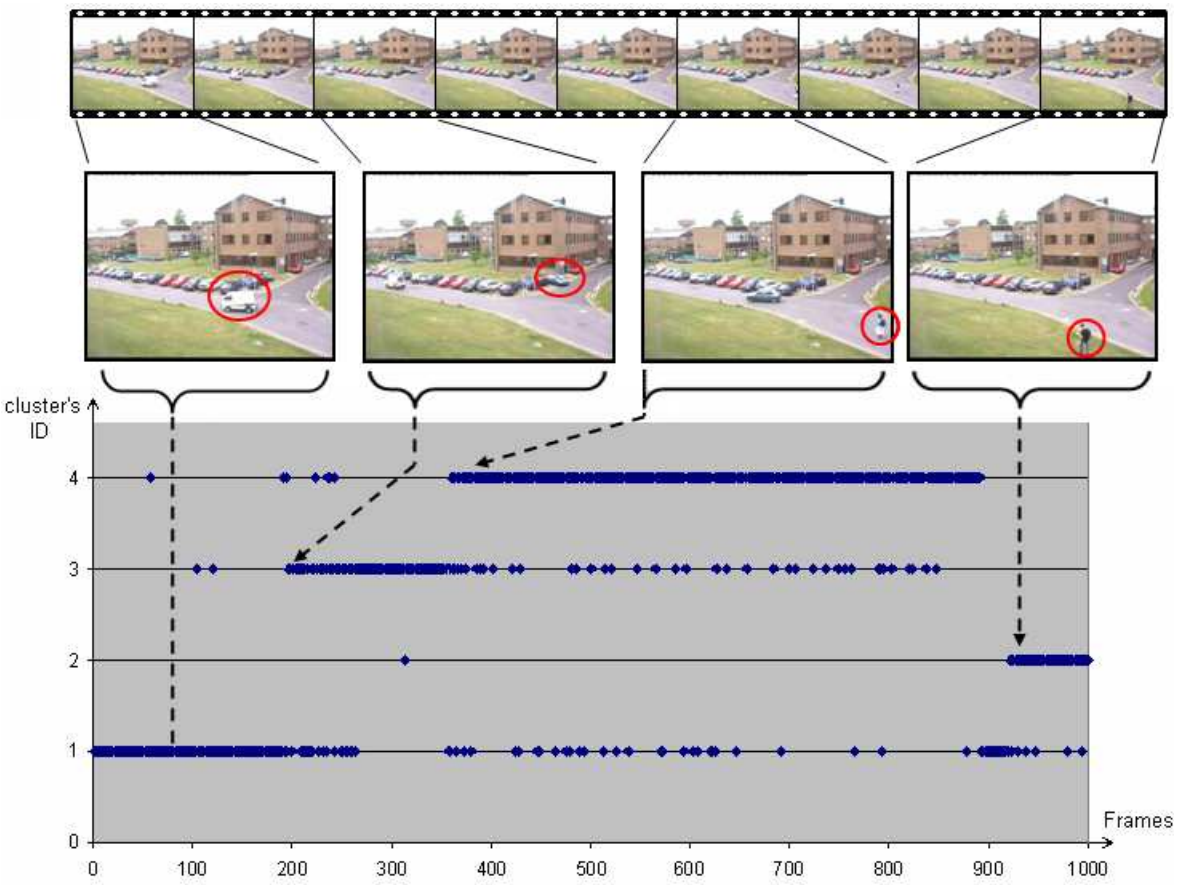

Figure 4. Clustering of frames from surveillance video.

the video by computing visual deviation of frames within a video segment. Simple threshold comparisons are used to identify static redundant segments. Visually similar scenes or multiple takes of the same scene are grouping together using our aforementioned approach. Basically, all video segments which do not contain the representative frame are classified as redundant. The results of the entire summarisation system can be found in [6]. Several criterions were used for summary evaluation including: IN - fraction of inclusions found in the summary (0 - 1); JU - summary contained lots of junk (1 strongly agree - 5 strongly disagree), RE - summary contained lots of duplicate video (1 strongly agree - 5 strongly disagree). Evaluation results of our method show following numbers: IN = 0.4, JU = 3.4 and RE = 4. Mean number of selected and redundant segments detected by our approach per a video is 5,3 and 65,5 respectively.

5. Conclusion

We have shown that Ant-Tree Strategy can be efficiently used for summarising of videos with the emphasis on the scene recognition. The approach generates a set of representative shots and extracts the tree structure of a video sequence. The tree structure generated by ATS is not strictly equivalent to a dendogram as used in standard hierarchical clustering techniques. Here each node corresponds to one data point, while in the case of dendograms data points correspond to leaves. Proposed method contributed to the automatic summarisation system by producing list of segments with very low duplication and redundancy. In our future work we will investigate the issue of intelligent combination of low-level features in order to decrease the semantic gap in video scene detection and efficient video summarisation.

References

[1] F. Azuaje and N. Bolshakova. Improving expression data mining through cluster validation. 2003.

[2] N. Azzag, H. Monmarch, M. Slimane, G. Venturini, and C. Guinot. Antree: a new model for clustering with artificial ants. *IEEE Congress on Evolutionary Computation*, pages 08–12, 2003.

[3] J. Calic and E. Izquierdo. Towards real time shot detection in the mpeg compressed domain. In *Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services*, 2001.

[4] A. Colorni, M. Dorigo, and V. Maniezzo. Distributed optimization by ant colonies. In *Proceedings of European Conference on Artificial Life*, pages 134–142, The Netherlands, 1991.

[5] U. Damnjanovic, T. Piatrik, D. Djordjevic, and E. Izquierdo. Video summarisation for surveillance and news domain. In *Proc. International conference on Semantics and digital Media Technologies*, 2007.

[6] E. Dumont, B. Mérialdo, S. Essid, W. Bailer, H. Rehatschek, D. Byrne, H. Bredin, N. E. O'Connor, G. J. Jones, A. F. Smeaton, M. Haller, A. Krutz, T. Sikora, and T. Piatrik. Rushes video summarization using a collaborative approach. In *TVS '08: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 90–94, Vancouver, British Columbia, Canada, 2008.

[7] R. Eberhart and K. J. A new optimizer using particle swarm theory. In *6th International Symposium on Micro Machine and Human Science*, pages 39–43, 1995.

[8] J. H. Holland. *Adaptation in natural and artificial systems, Ann Arbor*. The University of Michigan Press, 1975.

[9] N. Labroche, N. Monmarche, and G. Venturini. Antclust: Ant clustering and web usage mining. In *Genetic and Evolutionary Computation Conference*, pages 25–36, Chicago, 2003.

[10] J. Lee, G. G. Lee, and W. Y. Kim. Automatic video summarizing tool using mpeg-7 descriptors for personal video recorder. *IEEE Transaction on Consumer Electronics*, vol. 49:742–749, 2003.

[11] Z. Li, G. Schuster, and A. K. Katsaggelos. Minmax optimal video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15:1245–1256, 2005.

[12] I. Osuka, R. Radharkishnan, M. Siracusa, A. Divakaran, and H. Mishima. An enhanced video summarization system using audio features for personal video recorder. *IEEE Transactions on Consumer Electronics*, Vol. 52:168–172, 2006.

[13] V. Ramos and V. Almeida. Artificial ant colonies in digital image habitats - a mass behavior effect study on pattern recognition. In *Proceedings of ANTS'2000 - 2nd International Workshop on Ant Algorithms*, pages 113–116, Brussels, Belgium, 2000.

[14] Z. Rasheed and M. Shan. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, Vol. 7:1097–1105, 2005.

[15] N. Slimane, N. Monmarche, and G. Venturini. Antclass: discovery of clusters in numeric data by an hybridization of an ant colony with kmeans algorithm. Technical Report 213, 1999.

[16] G. Theraulaz, E. Bonabeau, C. Sauwens, J. L. Deneubourg, A. Lioni, F. Libert, L. Passera, and R. V. Sole. Model of droplet formation and dynamics in the argentine ant (linepithema humile mayr). *Bulletin of Mathematical Biology*, 2001.

[17] E. Venequ and et al. From video shot clustering to sequence segmentation. *IEEE International Conference on Pattern Recognition*, pages 254–257, 2000.

[18] Y. Wang, T. Zhang, and D. Tretter. Real time motion analysis towards semantic understanding of video content. In *Conference on Visual Communications and Image Processing*, 2005.

[19] L. B. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5:533–544, 1995.