A Graph Based Approach to Speaker Retrieval in Talk Show Videos with Transcript-Based Supervision*

Yina Han[1,**], Guizhong Liu[1,***], Hichem Sahbi[2], and Gérard Chollet[2]

[1] The School of Electronic and Information Engineering,
Xi'an Jiaotong University, 710049, Xi'an, China
ynhan@mailst.xjtu.edu.cn, liugz@xjtu.edu.cn
[2] CNRS LTCI UMR 5141, TELECOM-ParisTech, 75634, Paris, France
{hichem.sahbi,gerard.chollet}@TELECOM-ParisTech.fr

Abstract. This paper proposes a graph based strategy to retrieve frames containing the queried speakers in talk show videos. Based on who is speaking and when information from the audio transcript, an initial audio-based step, that restricts the queried person to frames corresponding to when he/she is speaking, with a second step that analyzes visual features of shots is combined. Specifically, based on the production property of talk show video, (1) Shot based graph is constructed first. Then the densest sub-graph is returned as the final result. But instead of direct search (DS) of the densest part, (2) We model the intra node connection and inter node connection by a frame layer degree map to take into account the duration information within each shot node; (3)A graph partition strategy without restriction on the shape and the number of sub-graphs is proposed, in which shots containing the same person are more similar to each other. Experiments on one episode of the French talk show "Le Grand Echiquier" show more than 10% improvement to audio only method and more than 7.5% improvement to DS method on average.

Keywords: Speaker retrieval, talk show video, multi-modality, graph.

1 Introduction

Person retrieval is essential to understand and retrieve real content from videos/images which are strongly related to human subjects [6, 7]. Recently, how to explore other available information, such as transcripts [1, 3] and captions [4,5,6,7,10] to facilitate the search work has been actively studied.

* This work is supported in part by the National 973 Project under Project No. 2007CB311002, and the National 863 Program under Project No. 2009AA01Z409. This material is based upon work funded by European K-Space Project and French Infom@gic Project.

** The author Yina Han would like to thank Prof. Henri Maitre for his valuable and constructive comments that helped improve the presentation of this paper.

*** Corresponding author.

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 962–967, 2009.

For talk show videos an essential clue for finding a speaker is when he/she is speaking. Hence, in this paper we first limit the search according to his/her speaking time recorded in the audio transcript. Then we propose a graph based strategy to analyze the visual features to refine the initial result set. Based on the following assumptions: (i) a speaker appears more frequently when he/she is speaking than other one else does; (ii) each speaker is presented by a set of relatively fixed scenes in our talk show video [8]; (iii) the shots presenting the same speaker tend to be visually more similar than the shots of different speakers. Hence a video shot serves as a basic structural unit to construct a graph. To emphasize the more frequently criteria in shot layer graph, we model the intra node connection and inter node connection by a frame layer degree map that improves retrieval results from direct search (DS) by taking into account the duration information within each node. To emphasize the more similar criteria in shot layer graph, we propose a graph partition strategy without restriction on the shape and the number of sub-graphs. In each separated sub-graph, shots containing the queried person are more similar.

The rest of this paper is organized as follows: First, the baseline algorithm, namely direct search of the densest part on shot layer graph is presented in Section 2. In Section 3 the proposed node duration and sub-graph partition strategy for shot layer graph is introduced. And experimental results and some conclusions are presented in Section 4 and Section 5 respectively.

2 Direct Search of the Densest Part on Shot Layer Graph

In this paper, RGB color histograms are used to represent the shots, and Chi statistic and histogram intersection (IS) are adopted to measure the shot similarity. Then a graph $G=(V,E)$ can be defined, where nodes in V represent shots and edges in E represent similarity between shots. To refine the initial results based on audio transcript (audio), we search for the densest sub-graph.

The density of subset S of a graph G is defined as [12]. $E(S)=\{(i,j)\in E: i\in S, j\in S\}$, namely the set of edges induced by subset S. The subset S that maximizes f(S) is defined as the densest component. $2f(S)$ is

$$f(S)=\frac{|E(S)|}{|S|} \tag{1}$$

simply the average degree of the sub-graph S, which is a discrete function, and its maximum solution is not unique. In [12] a greedy algorithm was proposed to find the densest sub-graph. The greedy search starts with the entire graph as subset ($S=V$). At each iteration, $f(S)$ is computed and the node with the minimum degree within S is removed. Finally, the subset with the highest encountered density is returned as the densest component of the graph.

3 Node Duration and Sub-graph Partition for Shot Layer Graph

The method above is subject to two severe problems: (i) the duration information within each shot is lost; (ii) each speaker is always presented by more than one type of shots which have no visual similarity.

3.1 Frame Layer Degree Map

To emphasize the impact of shot duration, we introduce a frame layer degree map. Given a shot layer graph $G=(V,E)$, a $|V|\times|V|$ adjacency matrix $A=\left[a_{i,j}\right]$, $i=1,2,\ldots,|V|;\ j=1,2,\ldots,|V|$ is used to store the structure, that is if there is an edge from vertex v_i to vertex v_j, the element $a_{i,j}$ is 1, otherwise it is 0. Note each shot node itself is a complete graph constructed by the frames involved, which we call it intra node connection; and an edge between shot node we call it inter node connection. We define a frame layer degree map $FD=\left[fd_{i,j}\right], i=1,2,\ldots,|V|;\ j=1,2,\ldots,|V|$ to describe these frame layer connections. According to the definition of degree for v_i complete graph, frame layer degree map is defined as:

$$fd_{i,j}=\begin{cases} f_i\left(f_i-1\right)/2 & \text{if } i=j \\ \left(f_i+f_j\right)\left(f_i+f_j-1\right)/2 & \text{if otherwise} \end{cases} \tag{2}$$

where f_i and f_j are the number of frames within node v_i and v_j respectively. Then the adjacency matrix A is modified as:

$$\tilde{A}=\left[a_{ij}\times fd_{ij}\right] \tag{3}$$

Finally, the greedy algorithm [12] is carried out on the adjacent matrix $\tilde{A}$ in order to find the densest sub-graph in frame layer.

3.2 Shot Graph Partition

To resolve the visual similarity problem, a graph partition method is proposed to separate varied shot appearances of the same person into different sub-graphs. Recalling the idea of mode seeking in [11]: given a set $\{x_i\}_{i=1\ldots n}$ of n points in the d dimensional space R^d, the multivariate kernel density estimate with kernel profile $k\left(\|x\|^2\right)$ and window radius (bandwidth) h, computed in the point x is given by

$$f(x)=\frac{C}{nh^d}\sum_{i=1}^{n}k\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \tag{4}$$

where C is a normalization constant to insure $f(x)$ is a probability density function. The modes of the density function are located at the zeroes of the gradient function $\partial f(x)/\partial x=0$. Denoting $g(x)=-k'(x)$, the gradient of (4) is:

$$\frac{\partial f(x)}{\partial x} = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) g\left(\left\| \frac{x - x_i}{h} \right\|^2 \right). \tag{5}$$

Given a graph $A = \left[a_{i,j} \right]$, regarding its nodes as points, its edges as distances, and when adopting the Epanechnikov profile, $g(x)$ is simplified as a constant [12]. Hence the gradient equation (5) for node c can be written as:

$$\frac{\partial f(c)}{\partial c} = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} a_{c,i} \ . \tag{6}$$

The mode that satisfies $\partial f(x) / \partial x = 0$ can be approximated by the node that has the minimum sum of distances to all the other nodes, namely:

$$c^* = \arg\min \left(\sum_{i=1}^{n} a_{c,i} \right). \tag{7}$$

Given a graph $G = (V, E)$, and bandwidth h, the specific partition process is described in the following steps:

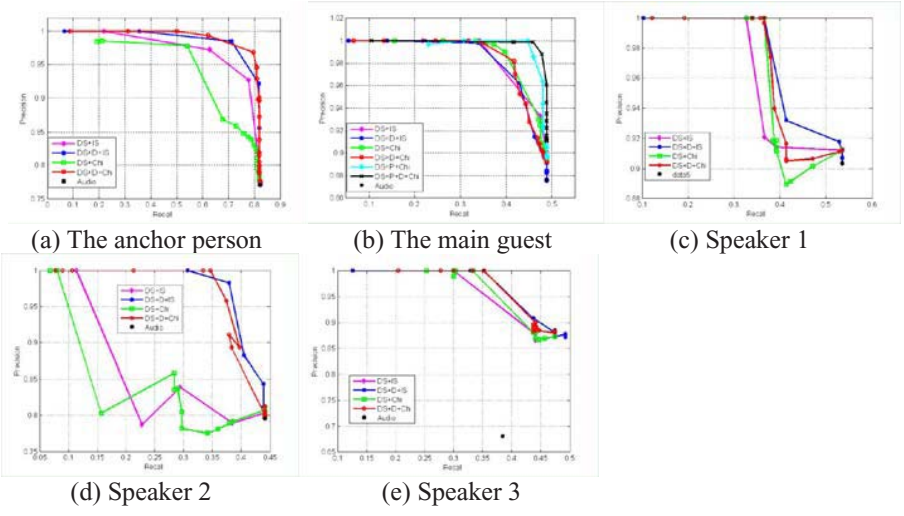
1. **Initialization:** Randomly select a node from V, for all nodes within bandwidth h find c_1 according to (7).
2. **Mode search:** Denote the above sub-graph as SG_1. For the remaining nodes in V, continue step 1, until no node is left.
3. **Grouping:** Delineate in the joint domain the sub-graph $\{SG\}_{p=1\ldots m}$ by grouping together all c_i which are closer than $h/2$.

Then we get sub-graphs without restriction on the shape or prior knowledge of the number of sub-graphs as shown in Fig. 2. The dense value for each sub-graph SG_i, namely $f(SG_i) = |E(SG_i)| / |SG_i|$, is calculated and those lower than a predefined threshold T, which are usually occasionally appeared false shots, are eliminated. Within each sub-graph, the shots of the same person are more similar than arbitrary shots. Finally, the greedy densest search algorithm is conducted for each sub-graph.

4 Experiments

The proposed method was applied to one episode of a French TV show "Le Grand Echiquier", which contains 1 hour and 43 minutes talking part between the anchor person (Jacques), the main guest (Gerard) and other related speakers.

The performance is evaluated by standard precision and recall criteria. Two baseline methods are compared to the proposed method: (i) audio transcript based method (denoted as *Audio*), and (ii) direct search on shot graph (denoted as *DS)* method.



(a) The anchor person (b) The main guest (c) Speaker 1

(d) Speaker 2 (e) Speaker 3

Fig. 1. Precision-recall curves, where DS means direct search on the shot graph, IS means taking histogram intersection as similarity measure, Chi means taking Chi statistics as similarity measure, D means introducing duration information by our proposed frame layer degree map, P means performing our proposed graph partition strategy before searching, and Audio means audio based initial search.

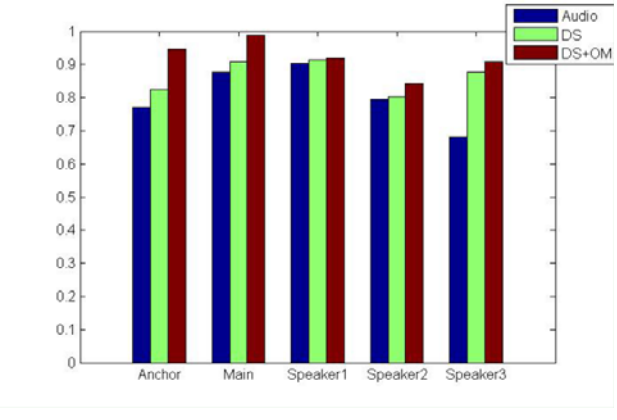


Fig. 2. Comparison of precision at the same recall rate, using the audio only method (Audio), direct search on the shot graph (DS), and our modification to direct search (DS+OM).

For the anchor person and other three speakers, we test two similarity measures, namely *Chi*statistics and histogram intersection (*IS*), as shown in Fig. 1.

The retrieval precision for the five speakers by using $Audio$, DS method and our modification (denoted as $DS{+}OM$) is compared in Fig. 2. Our method performs more than 10% improvement on average to $Audio$ only method and more than 5% improvement on average to DS method. That is up to 17.54% improvement to $Audio$

only method and up to 12.20% improvement to DS method for the anchor person; and up to 11.24% improvement to $Audio$ only method and up to 7.84% improvement to DS method for the main guest.

5 Conclusions

In this paper, we propose a novel graph based method to remove false alarms from audio based initial search space for talk show videos. Experimental results on one episode of the French TV show "Le Grand Echiquier" show more than 10% precision improvement to audio only method and more than 7.5% precision improvement to direct densest sub-graph search method for five main speakers. However, the recall rate is limited by the audio based initial search, that means frames without corresponding speech will never be retrieved. Our future work will focus on setting up appropriate association model between audio and video to improve recall rate.

References

1. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is Buffy"- Automatic naming of characters in TV video. In: BMVC (2006)
2. Sivic, J., Everingham, M., Zisserman, A.: Person spotting: video shot retrieval for face sets. In: ACM CIVR, pp. 226–236 (2005)
3. Sivic, J., Everingham, M., Zisserman, A.: Who are you? – Learning person specific classifiers from video. In: IEEE CVPR (2009)
4. Ozkan, D., Duyqulu, P.: A graph based approach for naming faces in news photos. In: IEEE CVPR, pp. 1477–1482 (2006)
5. Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Yee-Whye, T., Learned-Miler, E., Forsyth, D.A.: Names and faces in the news. In: IEEE CVPR, pp. 848–854 (2004)
6. Satoh, S., Kanade, T.: Name-It: Association of face and name in Video. In: IEEE CVPR, pp. 368–373 (1997)
7. Yang, J., Chen, M.Y., Hauptmann, A.: Finding person x: Correlating names with visual appearances. In: ACM CIVR, pp. 270–278 (2004)
8. Han, Y., Liu, G., Chollet, G., Razik, J.: Person identity clustering in TV show videos. In: IET VIE, pp. 488–493 (2008)
9. Han, Y., Razik, J., Chollet, G., Liu, G.: Speaker Retrieval for TV Show Videos by Associating Audio Speaker Recognition Result to Visual Faces. In: Proceedings of the 2nd K-Space PhD Jamboree Workshop (2008)
10. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Automatic Face Naming with Caption-based Supervision. In: IEEE CVPR, pp. 1–8 (2008)
11. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. PAMI. 24, 603–619 (2002)
12. Charikar, M.: Greedy approximation algorithms for finding dense components in a graph. In: APPROX: Int. Workshop on Approximation Algorithms for Combinatorial Optimization (2000)