

A New Anchor Shot Detection System for News Video Indexing

Hansung Lee¹, Younghee Im¹, Jooyoung Park², Daihee Park¹

¹ Dept. of Computer and Information Science, Korea University
E-mail: {mohan, yheem, dhpark}@korea.ac.kr

² Dept. of Control and Instrumentation Engineering, Korea University
E-mail: parkj@korea.ac.kr

ABSTRACT

In this paper, we present a new anchor shot detection system which is a core step of the preprocessing process for the news video analysis. The proposed system is composed of four modules and operates sequentially: 1) skin color detection module for reducing the candidate face regions; 2) face detection module for finding the key-frames with a facial data; 3) vector representation module for the key-frame images using a non-negative matrix factorization; 4) anchor shot detection module using a support vector data description. According to our computer experiments, the proposed system shows not only the comparable accuracy to the recent other results, but also more faster detection rate than others.

Key Words : Anchor Shot Detection, SVM, SVDD, NMF

1. Introduction

Nowadays, video materials and video services have been more available than ever before with respect to the information repositories as well as the commercial perspectives. Especially, according to the recent literature reviews, the news videos happened to be attracted by many researchers for the purpose of the news video analysis such as news video indexing, browsing, and retrieval. In fact, recently the news videos appeared to be very valuable information for the data analysts, information providers and TV consumers because of their information richness[1].

In a preprocessing process of the news video analysis, Anchor Shot Detection(ASD) is a core step for news video story segmentation. In literatures, there are two main research paradigms for ASD strategies in general[2-3]: 1) template matching method and unsupervised method. Template based method defines a model in advance for an anchor shot and match it against all the

shots of a news video. This method is proven to be too sensitive to the conditions of the studio as well as positions of anchor person. Consequently, it turned out to be not cost-effective. On the other hand, unsupervised method is grouping shots with similar visual contents that repeatedly occurred in the whole news video. Since this method consists of two phases which require the computational cost: clustering and pruning, it might be computationally expensive. Besides, there are also ongoing attempts to use audio-visual features for ASD[4-5]. However these methods are somewhat time-consuming from the practical point of view. Since news videos are produced every day with a great volume, it is necessary to devise the fast and accurate algorithms for ASD, needless to say.

In this paper, we introduce our novel ASD system which consists of four modules and operates sequentially: 1) skin color detection module for reducing the candidate face regions; 2) face detection module for finding the key-frames with a facial data; 3) vector

representation module for the key-frame images using a non-negative matrix factorization; 4) anchor shot detection module using a support vector data description.

2. New Anchor Shot Detection System

In this section, we introduce our newly proposed anchor shot detection system, which is given in Figure 1, aiming at the fast detection as well as high precision rate. Each module will be described one by one in details.

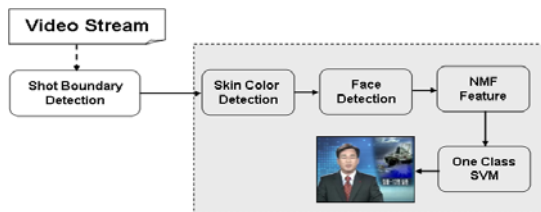


Fig. 1 The Architecture of ASD System

2.1 Skin Color Detection Module

To choose the candidate sets of anchor shots and reduce the search space for the real-time face detection phase, we adopt a fast skin color detection algorithm[6] in this stage which eliminates key-frames of each shot with no region of skin color or too large region. Now we employ a simple but cost-effective heuristic rule defined in (1). Consequently, the reduced image with a skin color region is generated with a comparable fast speed.

$$\begin{aligned}
 (R, G, B) \text{ is classified as skin if: } & \quad (1) \\
 R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \\
 R - G > 15 \text{ and } R > B
 \end{aligned}$$

2.2 Real-Time Face Detection Module

A real-time face detection module classifies the reduced images of each key-frame produced at the prior step into one with a facial data and the other with non-face data. Therefore, if the face is not detected or over 3-faces are detected in a key-frame, it is discarded automatically. Accordingly, we can reduce the candidate sets of anchor shot.

To meet our design requirements (i.e., speed and precision), we choose a real-time face detection model[6] which consists of several weak classifiers in a way of cascading structure and a support vector machine(SVM) in the last stage. A series of weak classifiers generates the candidate faces fast with low precision. Therefore, they are able to decide the candidate faces with low computational cost since the false-negative ratio is high but the false-positive ratio is close to zero. On the other hand, a SVM in the last stage detects a face from candidate faces. The architecture of real-time face detection module is given in Figure 2.

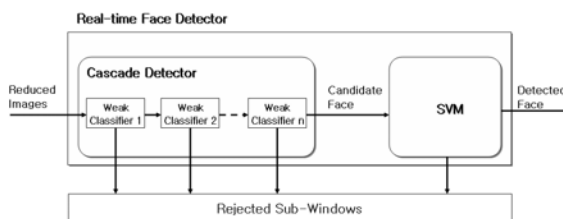


Fig. 2 The Architecture of Real-Time Face Detection System

2.3 Vector Representation Module

For the vector representation of key-frame images, we adopt a non-negative matrix factorization(NMF), since it has a distinctive characteristic comparing to other linear algebra techniques such as a singular value decomposition(SVD) and a principal component analysis(PCA). That is, in the induced space derived by a NMF, each axis captures the special features of each class, and each image is represented with an additive combination of axes. Each axis in the space derived by the NMF has more straightforward correspondence with each image class than in the space derived by SVD and PCA[7].

The NMF factoring a given positive matrix subject to positive constraints is derived as follows[7]: The image corpus is regarded as an $n \times m$ matrix V , each column contains n non-negative pixel values of one of the m images. Given matrix V , the NMF is defined as a problem to find non-negative matrix factors W and H such that:

$$V \approx WH \quad (2)$$

where $W = [w_{ij}]$, $H = [h_{ij}]$, $0 \leq i \leq n$, $0 \leq j \leq m$.

It is formalized through the following optimization problem.

$$\begin{aligned} \min \quad & J = \frac{1}{2} \|V - WH\|^2 \\ \text{s.t.} \quad & W, H \geq 0. \end{aligned} \quad (3)$$

Using the Lagrange function and the Kuhn-Tucker condition, the following equations for w_{ij} and h_{ij} are given:

$$(VH)_{ij}w_{ij} - (WH^TH)_{ij}w_{ij} = 0 \quad (4)$$

$$(V^TW)_{ij}h_{ij} - (HW^TW)_{ij}h_{ij} = 0 \quad (5)$$

These equations lead to following update formulas:

$$w_{ij} \leftarrow w_{ij} \frac{(VH)_{ij}}{(WH^TH)_{ij}} \quad (6)$$

$$h_{ij} \leftarrow h_{ij} \frac{(V^TW)_{ij}}{(HW^TW)_{ij}} \quad (7)$$

For the vector representation of image, NMF is usually defined as follows:

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{\alpha=1}^r W_{i\alpha}H_{\alpha\mu} \quad (8)$$

The r columns of W are called basis. Each column of H is called an encoding and is in one-to-one correspondence with a image in V . The rank r of factorization is generally chosen so that $(n+m)r < nm$, and the product WH can be regarded as a compressed form of the data in V .

2.4 Anchor Shot Detection Module

In the last stage of our system, we classify the anchor shot using a SVM for high precision. In general, the volume of data necessary for training varies depending on each class anchor shot data and non-anchor shot data. Hence, the training result may be distorted by other class owing to the unbalanced size of training data. Accordingly, it is preferable to select the decision boundary using one-class SVM (one of the most well-known one-class SVM is a support vector data description

(SVDD)). A SVDD is derived as follows[8]:

Given a dataset of N -patterns in d -dimensional input space, $D = \{\mathbf{x}_i \in \mathbb{R}^d \mid i = 1, \dots, N\}$, one-class SVM is defined as a problem to obtain a sphere that minimizes the volume of it including the training data, and it is formalized through the following optimization problem:

$$\begin{aligned} \min \quad & L_0(R^2, \mathbf{a}, \xi) = R^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall i. \end{aligned} \quad (9)$$

where \mathbf{a} is the center of the sphere that expresses a class, R^2 is the square value of sphere radius, ξ_i is the penalty term that shows how far i -th training data \mathbf{x}_i is deviated from the sphere, and C is the trade-off constant. By introducing a Lagrange function and saddle point condition, we obtain the following dual problem:

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^N \alpha_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i = 1, \quad \alpha_i \in [0, C], \quad \forall i \end{aligned} \quad (10)$$

A sphere can express more complex decision boundary in the feature space F and we can map an input space into a feature space using kernel function K . When the Gaussian function is chosen for the kernel, the problem can be further simplified as follows:

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i = 1, \quad \alpha_i \in [0, C], \quad \forall i. \end{aligned} \quad (11)$$

Note that in this case, the decision function of each class can be summarized as follows

$$\begin{aligned} f(\mathbf{x}) = & R^2 - (1 - 2 \sum_{i=1}^N \alpha_i K(\mathbf{x}_i, \mathbf{x})) \\ & + \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \end{aligned} \quad (12)$$

3. Experimental Results

To evaluate the effectiveness of a proposed anchor shot detection system, we collect a dataset from the two main Korean broadcasting stations, namely, KBS and MBC. A dataset is described in more details in Table 1. The ground truth is manually labeled in advance. We use the standard *precision* and *recall* criteria for evaluation measure.

Table 1 Dataset for evaluation

	# of Shots
Total Shot	1,226
Anchor Shot	55
Reporter	43
Interview	70

We obtained 399(32.5 %) and 148(12.1 %) key-frames as the candidate set without missing any anchor shots, respectively after a skin color detection and a face detection phase. By adopting a preprocessing process with a skin color detection and a face detection, we can eliminate the huge amount of non-candidate anchor shots effectively. As can be seen in Figure 3, only 10% shots including real anchor shots out of candidate sets was chosen after preprocessing. Therefore, we are entitled to claim that our system is very cost-effective.

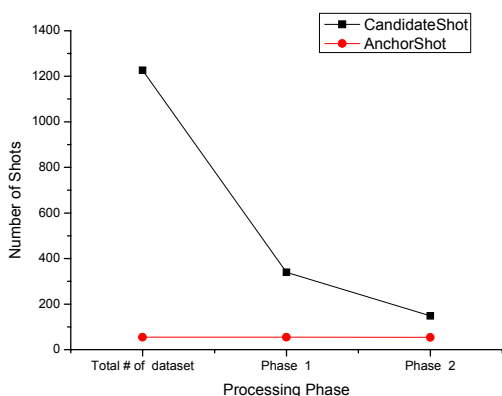


Fig.3 The # of candidate sets at each phase

Roughly comparing ours with the results of other previous methods[1-5] (in fact, it is not meaningful, since each method uses its own dataset, respectively), our system showed 98.14 % of precision and 96.8 % of recall, respectively which is comparable to others in

general.

4. Conclusions

In this paper, we proposed a novel anchor shot detection system which can detect anchor shot rapidly by reducing the search space. According to our computer experiments, the proposed system showed not only the comparable accuracy to the recent other results, but also more faster detection rate than others.

References

- [1] G. Xinbo, L. Jie, and Y. Bing, "A Graph-Theoretical Clustering based Anchorperson Shot Detection for News Video Indexing", *Proc. of ICCIMA'03*, pp. 108-113, 2003.
- [2] X. Luan, Y. Xie, L. Wu, J. Wen, and S. Lao, "AnchorClu: An Anchorperson Shot Detection Method Based on Clustering", *Proc. of PDCAT'05*, pp. 840-844, 2005.
- [3] M. Santo, P. Foggia, C. Sansone, G. Percannella, and M. Vento, "An Unsupervised Algorithm for Anchor Shot Detection", *Proc. of ICPR'06*, Vol. 2, pp. 1238-1241, 2006.
- [4] D. Lan, Y. Ma, and H. Zhang, "Multi-level Anchorperson Detection Using Multimodal Association", *Proc. of ICPR'04*, Vol. 3, pp. 890-893, 2004.
- [5] L. D'Anna, G. Marrazzo, G. Percannella, C. Sansone, and M. Vento, "A Multi-stage Approach for Anchor Shot Detection", *LNCS*, Vol. 4109, pp. 773-782, 2006.
- [6] J. Song, H. Lee, and D. Park, "Real-Time Face Detection System using Cascade Structure and SVDD", *Proc. of KCC05*, Vol. 32, No. 1(B), pp. 763-765, 2005.
- [7] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization", *Proc. of ACM SIGIR03*, pp. 267-273, 2003.
- [8] D. Tax and R. Duin, "Uniform Object Generation for Optimizing One-class Classifiers", *Journal of Machine Learning Research*, Vol. 2, Issue 2, pp. 155-173, 2001.