DATA-DRIVEN FUSION OF MULTI-CAMERA VIDEO SEQUENCES: APPLICATION TO ABANDONED OBJECT DETECTION

Suchita Bhinge, Yuri Levin-Schwartz and Tülay Adalı

Department of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County, MD, 21250

ABSTRACT

Due to the potential for object occlusion in crowded areas, the use of multiple cameras for video surveillance has prevailed over the use of a single camera. This has motivated the development of a number of techniques to analyze such multi-camera video sequences. However, most of these techniques require a camera calibration step, which is cumbersome and must be done for every new configuration. Additionally, these techniques fail to exploit the complementary information across these multiple datasets. We propose a data-driven solution to the problem by making use of the inherent similarity of temporal signatures of objects across video sequences. We introduce an effective solution for the detection of abandoned objects using this inherent diversity based on the transposed independent vector analysis (tIVA) model. By taking advantage of the similarity across multiple cameras, the new technique does not require any calibration and thus can be readily applied to any camera configuration. We demonstrate the superior performance of our technique over the single camera-based method using the PETS 2006 dataset.

Index Terms— Abandoned objects, joint blind source separation, multiple cameras, object detection, video surveillance

1. INTRODUCTION

Video surveillance is an active research field in computer vision. The aim of video surveillance is to efficiently extract useful information from large number of videos for object detection, tracking and activity recognition. The use of multiple cameras for surveillance has become popular since multi-camera sequences can view the monitored area from many angles, which reduces the issues of occlusion and limited visibility compared to single camera sequences. A number of techniques have been proposed to detect abandoned objects (AOs) from videos [1, 2, 3, 4]. These techniques, however, are based on the single camera paradigm and thus the performance deteriorates in complicated environments, such as with occlusion and crowd.

A number of smart surveillance techniques have been proposed to analyze multi-camera video streams [5, 6, 7, 8]. However, these systems first identify the topology of the cameras in order to connect the camera views by performing camera calibration, which must be done for every new configuration and hence is inefficient. Additionally, they do not exploit the complementary information across multiple cameras.

In order to exploit the diversity du to the statistically dependent temporal signatures of objects across cameras, in this paper, we propose a data-driven method to detect AOs based on transposed independent vector analysis (tIVA). The proposed method does not require the use of preprocessing stages such as camera calibration and topology estimation in order to link different camera views and does not depend on features such as color, size and position of the object as in [2, 8]. This means that the objects can have a different shape, color and size when viewed from different camera angles. Additionally, the proposed method has the potential to detect the AO even if it is not through all of the cameras and missing from a subset of those [9].

This paper is organized as follows. In Section 2, we discuss source separation using IVA, tIVA and order selection. In Section 4, we discuss the tIVA model for detecting abandoned objects and the detection technique proposed in this paper. We present our results in Section 5 using the PETS 2006 dataset [10] and compare our method with spatial independent component analysis (sICA) performed on each camera view. Finally, we discuss our results in Section 6.

2. BACKGROUND

2.1. IVA model

The general IVA model is given by $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}, k = 1, \ldots, K$, where the rows of $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times P}$ are the observations, $\mathbf{A}^{[k]} \in \mathbb{R}^{N \times M}$ is the mixing matrix and the rows of $\mathbf{S}^{[k]} \in \mathbb{R}^{M \times P}$ are the latent sources for the kth dataset. In order to estimate the latent source vectors $\mathbf{y}_m^{[k]}, m = 1, \ldots, M$, IVA estimates a demixing matrix, $\mathbf{W}^{[k]} \in \mathbb{R}^{M \times M}$, by minimizing the IVA cost function

$$\mathcal{I}_{IVA}(\mathcal{W}) = \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} \mathcal{H}\left\{ y_n^{[k]} \right\} - \mathcal{I}\left\{ \mathbf{y}_n \right\} \right\} - \sum_{k=1}^{K} \log |\det\left(\mathbf{W}^{[k]} \right)| \tag{1}$$

The representation in (1) explains that minimization of the cost function simultaneously minimizes the source entropy $\mathcal{H}\left\{y_n^{[k]}\right\}$ and maximizes the mutual information of nth source component vector (SCV), $\mathcal{I}\left\{\mathbf{y}_n\right\}$. The nth SCV is defined by concatenating the nth source from each dataset. The mutual information term in the IVA cost function plays an important role in exploiting the complementary information across all the datasets, since it maximizes the dependence within the nth SCV. Without this term, the IVA cost function would be equivalent to performing ICA on each dataset separately [11].

There are a number of IVA algorithms developed based on the assumption of the latent source distribution. IVA-Gaussian (IVA-G) [12], assumes the sources to be multivariate Gaussian and makes full use of second order statistics (SOS). IVA-Laplacian (IVA-L) [13] assumes a multivariate Laplacian distribution as the source prior and makes use of higher order statistics (HOS). IVA for multivariate generalized Gaussian distribution (IVA-GGD) [14] has a more general model than IVA-L and assumes a multivariate generalized Gaussian distribution (MGGD) as the source prior and thus can exploit both SOS and HOS. The MGGD covers a wide range of unimodal distributions such as super-Gaussian ($\beta < 1$), normal ($\beta = 1$) and sub-Gaussian ($\beta > 1$), where β is the shape parameter. In this paper, we use the IVA-GGD algorithm in order to exploit multiple types of diversity—SOS, HOS and source dependence.

2.2. IVA model for videos

An application of independent vector analysis (IVA) to video sequences has been introduced recently [15], in which the red, green and blue (RGB) color channels of the video frames form three datasets, $\mathbf{X}^{[k]}, k = 1, 2, 3$. Each dataset is defined as N observations by P samples, where N is the number of frames and P is the number of pixels. The columns of the mixing matrices describe the temporal behavior of the corresponding spatial component and is referred to as the time profile. The sources are time-independent components (foreground objects) and are estimated by exploiting spatial dependence across the RGB color channels. This model is, however, not suitable for the multi-camera paradigm. Video sequences from multiple cameras are shot from different angles in order to cover areas that may be occluded from some angles. These videos usually have common and distinct areas across different cameras and are therefore not spatially correlated, making the model used in [15] not a suitable match for multi-camera video sequences.

Though there is no spatial correlation, multi-camera sequences have a common time dimension since multiview videos are typically synchronized. Hence, if an object moves at time instant n in camera view 1, it moves at the same time in all other camera views that can see the object. Thus, in order to take advantage of the similarities in the temporal dimension, we propose a simple but powerful alternative through the transposition of the generative IVA model, the transposed IVA (tIVA) model, in order to detect AOs, which we introduce in the next section.

3. TRANSPOSED-IVA MODEL

TIVA is a modified representation of IVA that can automatically link the objects from multiple cameras based on their temporal signature. The tIVA model is given by $(\mathbf{X}^{[k]})^T = (\mathbf{S}^{[k]})^T(\mathbf{A}^{[k]})^T$. By transposing the datasets the role of observations and samples is reversed, *i.e.*, the sources are independent time profiles and the columns of the mixing matrix are the spatial components. Using this representation, the tIVA model estimates temporal signatures that are similar across multi-camera views. Hereafter, we denote the time profiles as $\mathbf{s}_m^{[k]}$ and the spatial components as $\mathbf{A}^{[k]}$. One advantage of the tIVA model is that the resolution of each camera can be different since we perform principal component analysis (PCA) on each dataset as described in the next section.

3.1. Order selection

Estimating the dimension of the signal subspace, *i.e.*, order selection, is an important issue since in many applications, the problem is overdetermined in nature and performing IVA on the signal subspace enable robust estimation of the components. For the tIVA model, the number of observations is equal to the number of pixels and the number of samples is equal to the number of frames. Since $N \ll P$, the maximum number of linearly independent signals must be less than or equal to N. The method in [16] provides a formulation for two information theoretic critera (ITC) under the independent and identically distributed (i.i.d.) sample assumption: Akaike's information criterion (AIC) and minimum description length (MDL). Since for videos, the samples exhibit pixel-wise dependence, this method can overestimate the order [17]. In [17] this issue is addressed by down-sampling the samples thus obtaining i.i.d. samples and applying the ITC formulation on the down-sampled dataset. Since order selection for the sample-poor regime is an open problem, we estimate the number of signals using the method proposed in [17] by using the regular model and use that order to perform PCA on the transposed model. The ITC formulation in [17] estimates eigenvalues of $\mathbf{X}\mathbf{X}^T$ using singular value decomposition, for which the singular values are square root of the eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$. Thus, the eigenvalues for the regular model and transposed model do not change, justifying the use of the order estimated using the regular model on the transposed model.

4. ABANDONED OBJECT DETECTION

Abandoned objects are defined as stationary objects that were not in the scene before. Based on this definition, the time profile of an abandoned object is expected to be a step-type response with an increase at the time instant when the object was placed. Thus, the tIVA model takes advantage of the correlation between the step-type responses across multi-camera

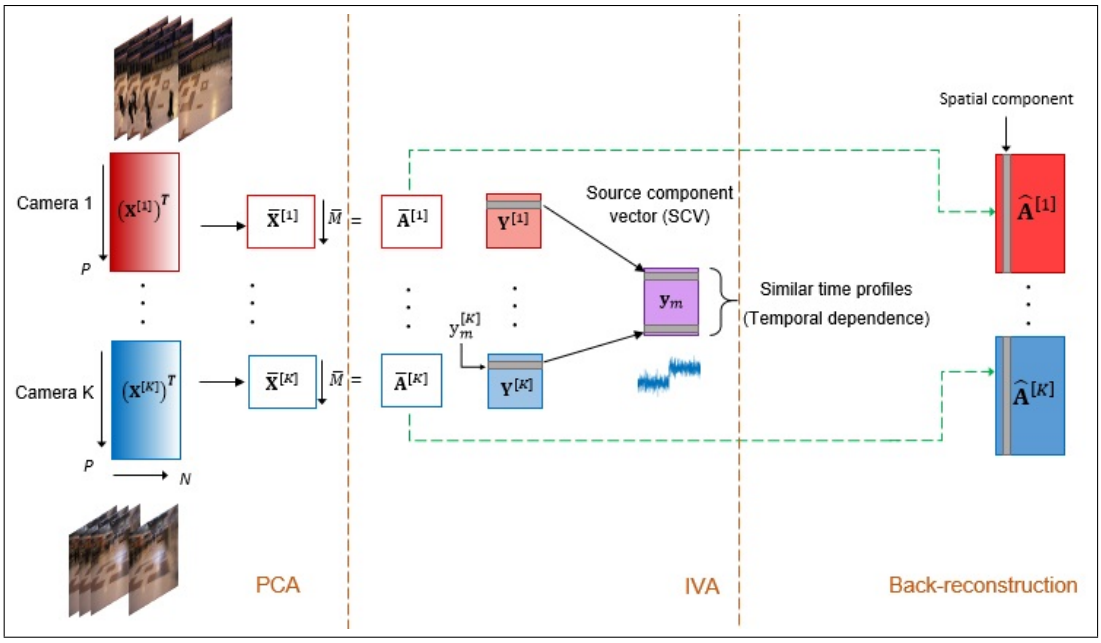

Fig. 1. Overview of implementation. Each dataset, $\mathbf{X}^{[k]}, k = 1, \ldots, K$ is defined as number of frames, N, by number of pixels, P. PCA is applied on each dataset, obtaining dimension reduced datasets, $\bar{\mathbf{X}}^{[k]}$. IVA-GGD is implemented on $\bar{\mathbf{X}}^{[k]}$ in order to estimate $\mathbf{W}^{[k]}$. The estimated sources, $\mathbf{Y}^{[k]}$, are the time profiles and the columns of the associated mixing matrix are the spatial components.

video sequences. In order to detect the AO, we look for a step-type response among the SCVs. Due to the complexity in the videos such as occlusion and crowd, the step-type responses are noisy, hence classical smoothing followed by gradient type techniques fails to detect the step [18]. For this reason, we implement a two-stage detection in which the first stage estimates the time point when a potential step occurred and the second stage determines whether the time profile is a step response or not.

In order to locate the point where the step change occurs, the time profile is correlated with an ideal step function and an area of interest is located that specifies the time points surrounding the step change. The length of the ideal step function is L time points, with $L/2$ time points before the step and $L/2$ time points after the step. The time points that pass an absolute correlation coefficient threshold, c_1, are labeled to be in the area of interest. Next, we perform a two-sample t-test at every time point in the area of interest in order to locate the exact time at which the step change occurred. This point is denoted as the estimated time of object drop, $\bar{T}$. In the second stage, we perform K-means clustering on the time profile in order to group the time points into $K = 2$ groups. A label vector, denoting the cluster to which each time point belongs, is obtained. This label vector is correlated with an ideal step response that has an increase at time point T and if the correlation coefficient is greater than c_2, the time profile is said to have a step-type response and the corresponding column of the mixing matrix shows the AO.

5. RESULTS AND DISCUSSION

We use the PETS 2006 dataset [10] to detect abandoned objects using the proposed tIVA model. The dataset consists of scenes of a train station recorded from 4 different angles. There are 7 scenarios, each scenario taking into account different challenges such as object occlusion and crowd. The data matrix $\mathcal{X} \in \mathbb{R}^{N \times P \times K}$ consists of four datasets, $K = 4$, N frames and $P = 576 \times 720 = 414,720$ pixels. Every 5th frame is picked since the frame rate is high (=25 frames/sec). Order selection is performed on each dataset separately as described in Section 3.1 and a median of all orders, denoted as $\bar{M}$, is used as the common order for all datasets. PCA is implemented on each dataset to form the feature matrix $\bar{\mathbf{X}}^{[k]} \in \mathbb{R}^{\bar{M} \times N}$ as shown in Figure 1. IVA-GGD is applied on $\bar{\mathbf{X}}^{[k]}$ in order to estimate $\mathbf{W}^{[k]}$. We choose β to be $0.4, 0.7, 4$ for this experiment in order to cover a wide range of unimodal distributions. The sources are estimated as $\mathbf{Y}^{[k]} = \mathbf{W}^{[k]}\bar{\mathbf{X}}^{[k]}$ and mixing matrices are back-reconstructed using $\hat{\mathbf{A}}^{[k]} = \left(\mathbf{F}^{[k]}\right)^{\dagger}\left(\mathbf{W}^{[k]}\right)^{-1}$, where $\mathbf{F}^{[k]}$ consists of eigenvectors corresponding to first $\bar{M}$ largest eigenvalues of $\left(\mathbf{X}^{[k]}\right)^{T}\mathbf{X}^{[k]}$ and $(\cdot)^{\dagger}$ denotes a pseudo-inverse operation. AO detection is performed using the rows of $\mathbf{Y}^{[k]}$ as described in Section 4 with $c_1 = 0.8, c_2 = 0.8$ and $L = 100$.

We demonstrate the advantage of the proposed method by comparing it with the results obtained by applying sICA on each dataset separately. The ICA algorithm used for comparison is entropy rate bound minimization (ERBM) [19, 20] which takes HOS and sample dependence into account as proposed in [21]. The performance is measured in terms of num-

Sequence	$\bar{M}$	Number of AOs	sICA (ERBM) CD/ID				tIVA CD/ID			
			V1	V2	V3	V4	V1	V2	V3	V4
S1	65	1	1/0	0/1	1/0	0/1	1/0	1/0	1/0	1/0
S2	52	1	0/1	0/1	0/1	0/1	1/0	1/0	1/0	1/0
S3	52	0	0/1	0/1	1/0	0/1	0/1	0/1	0/1	0/1
S4	58	1	1/1	0/1	0/1	1/0	1/0	0/1	1/0	0/1
S5	64	1	0/1	0/1	0/1	0/0	1/1	1/1	1/1	1/1
S6	60	1	0/1	0/1	0/1	0/1	1/0	1/0	1/0	1/0
S7	52	1	0/1	0/1	1/0	0/1	0/1	0/1	0/1	0/1
		6/0	2/6	0/7	1/5	3/4	5/3	4/4	5/3	4/4

Table 1. Comparison of tIVA and sICA in terms of number of correct detection (CD) and number of incorrect detections (ID). Red indicates incorrect results.

Sequence	Ground-truth T	sICA (ERBM) $\bar{T}$				tIVA $\bar{T}$			
		V1	V2	V3	V4	V1	V2	V3	V4
S1	375	376	447	366	416	374	376	367	391
S2	240	212	176	339	342	221	222	230	231
S3	-	367	375	-	268	165	167	168	163
S4	160	154	343	502	145	156	194	156	205
S5	360	532	261	533	-	353	352	356	357
S6	320	242	273	175	390	314	313	319	314
S7	210	234	356	209	263	264	264	264	264

Table 2. Comparison of tIVA and sICA with respect to the estimation of time of the bag-drop. Red indicates incorrect detections, i.e., not in the range $T - 20 \leq \bar{T} \leq T + 20$.

ber of correct detections (CD) and number of incorrect detections (ID) as shown in Table 1, and in terms of the time point of bag-drop, T as shown in Table 2. The detected component is correctly classified if the estimated time point $\bar{T}$ of bag-drop is close to the ground-truth, T, by $\pm$ 20 frames (4 seconds) and by visually looking at the spatial component (corresponding column of the mixing matrix).

From Table 2, we note that tIVA solves the issue of occlusion, *e.g.*, in scenario 2, the object is occluded by the people for some time when viewed from camera 3 and 4. Hence the estimated time point from sICA for $V3$ and $V4$ is not close to the ground-truth and is equal to the time point when the people moved away from the object. However, this issue is not observed in tIVA since it jointly estimates the time profiles. Note that tIVA was able to detect the AO in scenarios 4, 5 and 6, however it has an incorrect detection in scenarios 4 and 5 because in this video, a person walks into the scene and waits till the end of the video. An object detection scheme for humans can be used as a post-processing step in order to remove this false positive.

In scenario 3, tIVA was not able to classify the object as true negative. In this scenario, the person temporally places the bag on the floor before picking it up while another person walks into the scene just after the person has picked the bag and stands still for some time before leaving. tIVA fails in this case since it combines the time profiles of two temporary stationary objects into one. Similarly, in scenario 7, tIVA model combines two time profiles into one since there are two stationary objects: one is a abandoned bag and the other is a person who is sitting in the beginning and leaves later, so the person acts as a removed object. The time profile for a removed object is also a step response with a step decrease, however due to sign ambiguity in IVA, the two time profiles are similar. Thus, the time point of the bag drop shown in Table 2, is equal to the time point when the person leaves. The issue of tIVA combining two time profiles in these scenarios might be due to the algorithm choice. A step-response has a bimodal distribution, however IVA-GGD assumes the sources to have a unimodal distribution. Thus, flexible IVA algorithms, such as the extension of ICA-entropy bound minimization [22] to IVA can be developed to address this issue.

6. CONCLUSION

In this paper, we introduce a new technique to detect abandoned objects from multi-camera views using the tIVA model that takes advantage of the correlation between the time profiles of the objects across multiple views. We test our technique on the PETS 2006 dataset that includes 7 scenarios with varying levels of complexity and compare it with ICA applied on each view separately. In general, the results indicate that tIVA performed better than sICA in terms of correctly detecting the AOs and estimating the time of object drop. The proposed method increases our confidence in the detection results since the test to detect AO is performed on multiple datasets in contrast to sICA, for which the result relies on a single test.

7. REFERENCES

[1] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, "Real time, online detection of abandoned objects in public areas," in *Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA).*, May 2006, pp. 3775–3780.

[2] J. Kim and D. Kim, "Accurate static region classification using multiple cues for ARO detection," *Signal Processing Letters*, vol. 21, no. 8, pp. 937–941, Aug. 2014.

[3] Y. Tian, R. Feris, H. Liu, A. Hampapur, and M.-T. Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews.*, vol. 41, no. 5, pp. 565–576, Sept 2011.

[4] F. Porikli, "Detection of temporarily static regions by processing video at different frame rates," in *IEEE Conference on Proc. Advanced Video and Signal Based Surveillance (AVSS).*, Sept. 2007, pp. 236–241.

[5] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[6] R. Pflugfelder and H. Bischof, "Localization and trajectory reconstruction in surveillance cameras with nonoverlapping views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 709–721, April 2010.

[7] T. T. Santos and C. H. Morimoto, "Multiple camera people detection and tracking using support integration," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 47–55, 2011.

[8] M. D. Beynon, D. J. Van Hook, M. Seibert, A. Peacock, and D. Dudgeon, "Detecting abandoned packages in a multi-camera video surveillance system," in *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS) 2003.*, pp. 221–228.

[9] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Two effective models based on ICA and IVA and their properties," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1478–1493, Sept 2015.

[10] "PETS 2006 dataset," http://www.cvg.reading.ac.uk/PETS2006/data.html.

[11] M. Anderson, G. S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, Sept 2014.

[12] M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: algorithms and performance analysis," *Signal Processing*, vol. 60, no. 4, pp. 1672–1683, 2012.

[13] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2006, pp. 165–172.

[14] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis, the Kotz distribution, and performance bounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3243–3247.

[15] S. Bhinge, Z. Boukouvalas, Y. Levin-Schwartz, and T. Adalı, "IVA for abandoned object detection: Exploiting dependence across color channels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2494–2498.

[16] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, vol. 33, no. 2, pp. 387–392, 1985.

[17] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Human brain mapping*, vol. 28, no. 11, pp. 1251–1266, 2007.

[18] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[19] X.-L. Li and T. Adalı, "Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010, pp. 1934–1937.

[20] G.-S. Fu, R. Phlypo, M. Anderson, X.-L. Li, and T. Adalı, "Blind source separation by entropy rate minimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4245–4255, Aug 2014.

[21] S. Bhinge, Y. Levin-Schwartz, G. S. Fu, B. Pesquet-Popescu, and T. Adalı, "A data-driven solution for abandoned object detection: Advantages of multiple types of diversity," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2015, pp. 1347–1351.

[22] X.-L. Li and T. Adalı, "Independent component analysis by entropy bound minimization," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5151–5164, Oct 2010.