Meaningful Automatic Video Demultiplexing with Unknown Number of Cameras, Contrast Changes, and Motion

J.L. Lisani L. Rudin P. Monasse J.M. Morel P. Yu

Cognitech Inc.
Pasadena, CA (USA)
lenny@cognitech.com

Abstract

This paper presents a software-based parameter-free method for the demultiplexing of a video stream [7] that is missing camera labeling information. The method is based on the observation that frames coming from the same input camera share some common characteristic features. These features are extracted from the input frames and grouped together according to statistical criteria. As a result of this grouping the number of different input sources in the video stream is inferred and it is possible to ascertain the source for each frame.

1. Introduction

It is customary in video security CCTV appplication to store data coming from different sources into the same video stream. In order to retrieve useful information from the stored data, this video must be demultiplexed (see Figure 1). The multiplexed video stream is recorded together with the camera number label, in order to enable hardware demultiplexing. This label information is lost when video is digitized and stored in the computer memory for image processing and analysis. Since frames from different cameras at different times get shuffled together, this digitized video is not amenable to human viewing in this form.

In [8] the concept of algorithm (software) based demultiplexing was introduced. In a general situation it is not possible to deduce the source of a frame from its position in the video stream. That is, we can not assume that frames from camera i are always at positions $i, i+k, i+2k, \cdots$, where k is some fixed parameter. Moreover, the number of video sources is generally unknown. Therefore the problem consists in taking a multiplexed video sequence, to automatically deduce the number of video inputs and to generate a separate video sequence for each input. Figure 1 illustrates the problem.

The obvious method is to exploit local differences between frames to decide whether or not two frames are grouped together as corresponding to the same camera view. This approach assumes "quiet zones" present in the camera views, where no change or motion takes place. The "quiet zones" based algorithms were later used by other video surveillance image processing software manufacturers. However the assumption of no change zones is extremely limiting, and unpractical. In particular, the "quiet zones" methods critically depend on arbitrary thresholds, which decide how to separate camera views. Thus any change in the same camera image frame could trigger the threshold, thus separating camera consistent frames. In [6] the first method which is invariant to arbitrary time-dependent image changes (i.e. illumination changes, object motion changes, etc.) was proposed. The change-invariant method in [6] minimizes an energy functional depending on a grouping error and the number of obtained groups. Both terms are balanced in the functional by a scale parameter. Thus the demultiplexing was accomplished by automatically segmenting the multiplexed video stream into the separate, camera consistent video frame sequences. The method in [6] assumed that the number of cameras is known, and no illumination difference threshold is assumed.

In order to eliminate this constraint, and to have more robust demultiplexing method, we take a different approach: we assume that some global features are common to all the frames coming from the same video source. The robustness of this new algorithm comes from the fact that we are able to separate camera streams not just by the pixel brightness, but with any other contrast-invariant imaging measurements, e.g. direction of the image gradient. Therefore, by plotting the histogram of each feature, computed for the whole video sequence, we expect to find as many groups of values (i.e. histogram modes) as video sources in the video stream. A simple demultiplexing algorithm would consist then in grouping together all the frames contributing to the same mode.

Plan of the paper. As it has been pointed out in the

0-7803-9385-6/05/$20.00 ©2005 IEEE

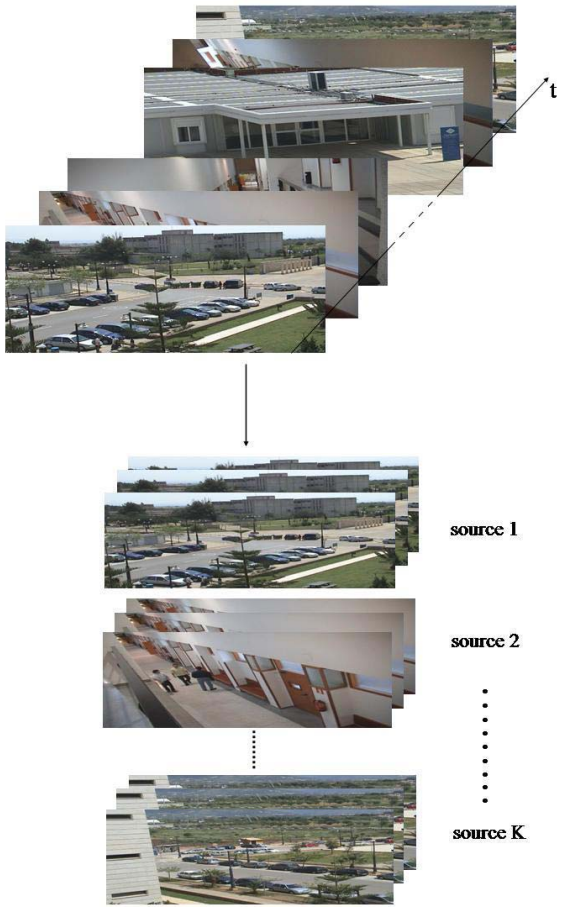

Figure 1: Video demultiplexing: images coming from different sources are stored in a single video stream which must be decomposed (demultiplexed) in as many video sequences as input sources.

previous paragraph, the proposed demultiplexing method strongly depends on the reliable extraction of modes in 1D histograms. In the next section a parameter-free method for modes extraction, based on the notion of "meaningfulness", is described. In Section 3 this method is applied to video demultiplexing. Section 4 displays some results of the application of the proposed demultiplexing algorithm and, finally, some conclusions are presented in Section 5.

2. Robust detection of modes in 1D-histograms

Roughly speaking, a "mode" in a 1D histogram can be described as an interval of values where data concentrate. In this sense, we say that the histogram in Figure 2 contains two modes. The precise definition of a mode is, however, not always easy and several criteria can be used. One of the most popular approaches assumes that the histogram is the result of a mixture of gaussian functions ([1]). Based on this assumption several methods have been proposed to compute the parameters (mean and variance) of the K gaussians that better fit the shape of the histogram, where K is a parameter that can be estimated by using different approaches ([2, 3]). It must be remarked that the method does not directly estimates the number, nor the location, of the modes in the histogram. Indeed, a mixture of K gaussians may contain any number of modes up to K. Moreover, in practice, not all the histograms can be modelled as a gaussian mixture.

Recently, a new method that automatically computes the number and the location of the modes of any 1D histogram has been described ([4]). The method is based on the notion of "meaningfulness": something is said to be "meaningful" if it can be considered as a deviation from randomness. The meaningfulness of an event (e.g. a geometric structure in an image, a mode of a histogram, etc.) is therefore measured as the result of a comparison to a random model (this is known as an *a contrario* method).

In the case of histograms the authors compare a given distribution of values (i.e. the actual histogram) with a random distribution, which they assume to be uniform. The following section summarizes the main results in [4], leading to the definition of the "meaningful modes" of a histogram.

2.1. Meaningful modes of a histogram

Let us consider a discrete histogram, that is, a finite number M of points distributed among a finite number L of values. We assume that the set of possible values is $\{1, \ldots, L\}$. For each discrete interval of values $[a, b]$, let $k(a, b)$ be the number of points with value in $[a, b]$, and let $p(a, b) = (b - a + 1)/L$. $p(a, b)$ represents the probability for a point to have a value in $[a, b]$ in a random (uniform) situation.

The probability that a fix interval $[a, b]$ contains at least $k(a, b)$ points among the M is $\mathcal{B}(M, k(a, b), p(a, b))$, where $\mathcal{B}(n, k, p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}$ denotes the tail of the binomial distribution of parameters n and p.

Since the total number of possible intervals in the histogram is $L(L+1)/2$, the expected number of intervals that contain at least $k(a, b)$ points among the M can be upper bounded by the quantity

$$NF_I([a, b]) = \frac{L(L+1)}{2} \mathcal{B}(M, k(a, b), p(a, b))$$

Definition 1. $[a, b]$ is a **meaningful interval** if $NF_I([a, b]) < 1$.

That is, an interval is meaningful if its expected number of occurrences in a random situation is less than one, which means that it contains more points than the expected average. Similarly, a "gap" can be defined as an interval that contains less points than the expected average.

Definition 2. $[a, b]$ is a **meaningful gap** if $NF_G([a, b]) < 1$, where

$$NF_G([a, b]) = \frac{L(L+1)}{2} \mathcal{B}(M, M - k(a, b), 1 - p(a, b))$$

Definition 3. $[a, b]$ is a **meaningful mode** if it is a meaningful interval and if it does not contain any meaningful gap.

Since a meaningful mode may contain or may be contained in another meaningful mode, an economy principle leads to the definition of maximal meaningful modes:

Definition 4. An interval A is a **maximal meaningful mode** if it is meaningful mode and if for all meaningful modes $B \subset A$, $NF_I(A) \leq NF_I(B)$ and for all meaningful modes $B \supsetneq A$, $NF_I(A) < NF_I(B)$.

It can be proved that two maximal meaningful modes never intersect.

Figure 2 shows an example of the detection of two maximal meaningful modes in a histogram.

A recent refinement of the method, which is able to find meaningful modes using as a random model a non-uniform distribution, has been proposed in [5].

3. Description of the method

As indicated in Section 1 the proposed method assumes that some global features are common to all the frames in the video stream coming from the same input source. Therefore, the analysis of the histograms associated to each one of these features should permit to infer the number of input sources and to associate every frame to a source.

The main difficulty of the approach is the definition of these global features. Two requirements must be fulfilled: each feature must be constant enough along all the frames from the same source to concentrate around a small set of values (a mode in the histogram); on the other hand, it must be significantly different between different video sources in order to obtain different modes for each source.

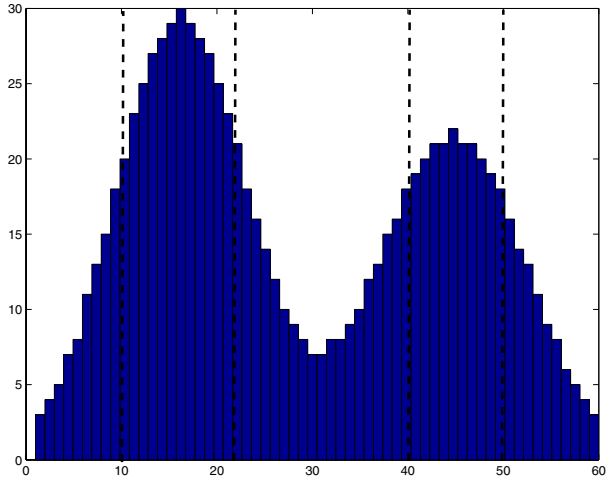

Figure 2: Example of a bi-modal (2 modes) histogram. The dashed lines indicate the location of the two meaningful modes of the histogram, computed by using the method described in [4].

As a first feature we propose the following quantity:

$$\bar{I}_i = \frac{1}{N} \sum_{(x,y)} I(x, y)$$

$\bar{I}_i$ is the mean illumination of frame i, N is the image size and $I(x, y)$ is the grey level at pixel (x, y). For RGB color images $I(x, y)$ is computed as the average of R, G and B.

Other global descriptors may be defined by the formula:

$$c_i^\lambda = \text{card}\mathcal{X}_\lambda^i = \text{card}\{(x, y)/I(x, y) \geq \lambda\}$$

that is, c_i^λ is the number of pixels in frame i having a grey level greater or equal to λ (in a digital image $\lambda \in \{0, \cdots, 255\}$).

Experimental results prove that either $\bar{I}_i$ as c_λ^i (for arbitrary values of λ) fulfill the first requirement, that is, they are almost constant for all the frames from the same video source. The reason is that objects moving in front of a CCTV camera use to be small compared to the background and, therefore, the illumination changes induced by them are negligible.

It is clear that frames coming from different video sources may have similar values of $\bar{I}$ or c_λ for some λ's. Nevertheless, if we define, for each frame, a vector of global features $v_i = (\bar{I}_i, c_{\lambda_1}^i, \cdots, c_{\lambda_n}^i)$, the probability that frames from different sources have similar vectors is very small, provided that the set of values $\lambda_1, \cdots, \lambda_n$ is large enough.

3.1. Demultiplexing algorithm

1. Select a set of n equally-spaced grey levels $\lambda_1, \cdots, \lambda_n$ in the range $[0, 255]$.

2. Define the following set of global features:

   $f_1 = \bar{I}, f_2 = c_{\lambda_1}, \cdots, f_{n+1} = c_{\lambda_n}$.

3. Set the feature index to $i = 1$.

4. Construct a histogram with the values of feature f_i computed for all the frames in the video sequence.

5. Compute the maximal meaningful modes of the histogram, using the definitions in Section 2.1.

6. For each mode in the previous histogram:

   i. Compute the sub-sequence composed by the frames that contribute to the mode.

   ii. For this sub-sequence, increase the feature index ($i \rightarrow i+1$) and repeat steps 4 to 6 just for the frames in the sub-sequence.

The sub-sequences obtained at the end of the algorithm contain the frames corresponding to each input source. The number of input sources is therefore equal to the number of obtained sub-sequences.

In practice, the number of values $\lambda_1, \cdots, \lambda_n$ doesn't need to be large. In fact, just two features, $\bar{I}_i$ and $c^i_{\bar{I}_i}$, have been used in our experiments, providing good results for all the tested video sequences.

Remark: quantization of the histograms. In order to compute the histograms of $\bar{I}$ and c_λ a quantization of these values is required. $\bar{I}$ values are quantized to their closest integers, which means that the data in the histogram of $\bar{I}$ are distributed among 256 values. Concerning c_λ, it takes integer values in the range $[0, N]$, where N is the image size. However, if the c^i_λ values are distributed in this range, even small variations in the images will affect the histogram, which might lead to the detection of several gaps and to an over-segmentation of the video stream. The quantization of this value must be based on the maximum allowed variation between frames from the same video source. A reasonable parameter is to consider that variations smaller than 1% or 0.5% of the image size are not significant enough and to quantize the range $[0, N]$ with 100 or 200 values.

4. Experimental Results

Figures 3 to 6 show an example of the application of the demultiplexing algorithm. The video stream in Figure 1 contains frames from 6 different sources. Its histogram of $\bar{I}_i$ values is composed of 4 meaningful modes (Figure 3). For each mode a video clip is obtained (Figure 4) and its histogram of $c^i_{\bar{I}_i}$ values is computed (Figure 5). Two of the histograms are uni-modal, which indicates that all the frames in the clip come from the same source. The other two are bi-modal, which means that the clips contain frames from two different sources. By grouping together the frames contributing to each mode, we obtain the six original video sequences that were multiplexed in the original video stream (Figure 6).

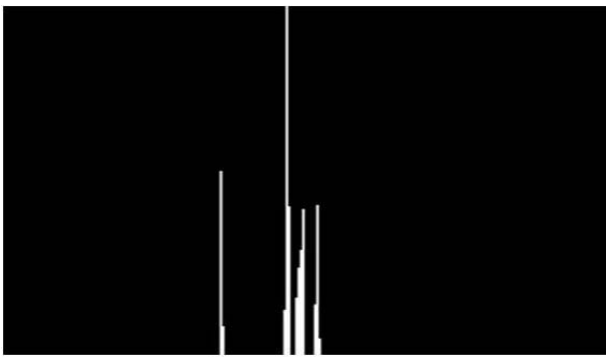

Figure 3: The video stream in Figure 1 contains frames from 6 different sources. Its histogram of $\bar{I}_i$ values is composed of 4 meaningful modes.

Figure 4: For each mode in the histogram of Figure 3 a video clip is obtained. This figure displays representative frames of the obtained clips. From left to right and from top to bottom the frames correspond to the four modes in the histogram (from left to right). Remark that some clips (top-right and bottom-left) contain frames coming from two different input sources.

5. Conclusions

The proposed demultiplexing algorithm has proven its performance with several multiplexed video streams. The algorithm can be applied offline, to the whole video sequence, or in a real-time environment, after a learning stage. In the later case, a large enough number of frames needs to be used in order to generate the histograms from which the global features characterizing each input source are learned. The

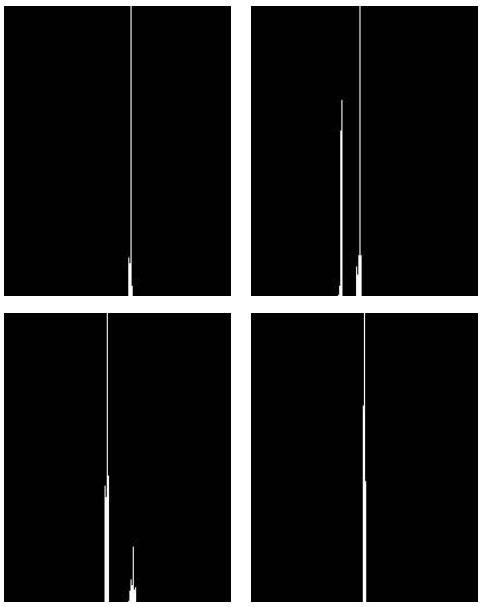

Figure 5: The histogram of $c^i_{\bar{I}_i}$ values is computed for each clip in the previous Figure. As expected, two of the histograms are uni-modal, which indicates that all the frames in the clip come from the same source. The other two are bi-modal, which means that the clips contain frames from two different sources.

Figure 6: Final result: frames from the six original video sequences that were multiplexed together in the video stream.

rest of the frames of the sequence are classified according to the values of their global features. A periodical update of the learned features is needed in order to take into account variations in illumination, background content, etc.

The demultiplexing tools based on the above discussed approach have been implemented in the commercial software tools for forensic video analysis and processing [9].

References

[1] D.M. Titterington, A.F.M. Smith, U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons Inc., New York, 1985

[2] A.B.M.L. Kabir, *Estimation of Parameters of a Finite Mixture of Distributions*, J. Royal Statist. Soc., Ser. B, vol. 30, pp. 472-482

[3] N.M. Nasab, M. Analoui, *Mixture Conditional Estimation Using Genetic Algorithms*, ISSPA, Kuala Lumpur, Malaysia, 2001.

[4] A. Desolneux, L. Moisan, J.M. Morel, *A groupinig principle and four applications*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, num. 4, pp. 508-513, 2003.

[5] J. Delon, A. Desolneux, J.L. Lisani, A.B. Petro, *Color Image Segmentation using Acceptable Histogram Segmentation*, accepted at IbPRIA05.

[6] F. Guichard, A. Litz, L. Rudin, P. Yu, *Software-based Universal Demultiplexing (Threshold-free Energy Minimization Approach)*, Proc. SPIE, vol. 4232, pp. 513-520, 2001.

[7] L. Rudin, J.L. Lisani, J.M. Morel, P. Yu, *Video Demultiplexing by Histogram Analysis Based on Meaningful Modes Extraction*, US Patent Office Application Filed, Cognitech, Inc., 2004.

[8] Cognitech Video Investigator 2000, Forensic video processing software © Cognitech 2000, Pasadena, CA.

[9] Cognitech Video Investigator, Cognitech Video Active 2005, Forensic video processing software © Cognitech 2000, Pasadena, CA.